Think Aloud, Non-Continuous Reporting, and Annotated Cloze

Using verbal report and a self-coding procedure in looking at German and Japanese informants' processing of a second language cloze task

Bob Gibson

Dissertation submitted August, 2005

I hereby declare that this thesis is entirely my own work.

Bob Gibson
August 15th, 2005
Wadamachi, Japan

Acknowledgments

I would like to extend my sincerest thanks to Prof. Alan Davies of (what used to be) Edinburgh University's Department of Applied Linguistics, and to Brian Parkinson of the University's Institute of Applied Language Studies. These advised well, criticised gently, and forbore beyond any call of duty.

My warm thanks, too, to the many individuals in Germany and Japan who acted as informants or consultants, or who helped with the tasks of translating and transcribing the data collected. These are too many to list here, but without the assistance of my wife, Shimada Yoshiko, I would have given up long ago. Jeff Hubbell of Hosei University played a mean sounding board. The surviving shortcomings of this thesis are due entirely to my own.

Bob Gibson

The freeware statistics program, Smith's Statistical `Package (SSP) for the Mac OS, used in some analyses here, was created by Gary Smith of Pomona College. Download the latest version from: http://economics.pomona.edu/StatSite.framepg.html

ABSTRACT

Following a survey of the history and nature of the cloze test as applied to second-language assessment, the study looks at the role of introspective and verbal-report data in understanding second-language test-taking processes. Particular attention is paid to the verbal report task format known as 'think-aloud'. The theoretical bases of this procedure are critiqued, and some problematic aspects of its use in the study of linguistic tasks such as cloze are discussed. Attention is drawn to apparent divergences between the standard model of think-aloud and its real-world applications.

The use of think-aloud in the study of cloze test-taking by German and Japanese first-language informants is discussed, and a number of specific shortcomings identified. These lie mainly in the areas of practical sample-size, interpretability and comprehensiveness of data, and negative affective responses on the part of informants.

A modification to the 'classical' think-aloud is proposed, labeled 'non-continuous reporting', and the results of this method are compared to those of think-aloud. It is concluded that the advantages of non-continuous reporting outweigh its shortcomings.

A further alternative real-time data-gathering procedure is proposed, the so-called 'annotated cloze', and its strengths and drawbacks discussed. The relative efficiencies of annotated cloze and the two variants of think aloud are examined in terms of their ability to generate a picture of how test-takers process cloze passages, and suggestions are offered regarding the use of these task formats in the elicitation of further data.

Table Of Contents

Chapter 8: Conclusions & suggestions for further research

Appendices

CHAPTER 1: INTRODUCTION AND RATIONALE

1.0 Introduction & rationale

Cloze is, for the overwhelming majority of those familiar with it today, a species of test. It is natural, therefore, that a thesis with 'cloze' in the title may be expected to consider the classical testing concerns of item facility, reliability, test-takers' scores, and the like. While I touch on some of these in later chapters, they are not my primary concern. In the following pages, I focus more on ways of understanding the processes and behaviours through which test-takers deal with react to test tasks, rather than on the facility and discriminative power of the test items themselves. This is in no way to denigrate these aspects of testing, which are in most practical contexts all that are really required in order to achieve the goals which testers typically set themselves and their instruments.

'Hard facts' in the form of facility indices, etc. can be gathered via the paper evidence test-takers leave behind, but the desire –increasingly widely felt, if one can judge from recent position papers and discussions in the testing literature– to know more about the routes by which test-takers produce this evidence calls for other methods. What I will try to set out here, following a review of some (and only some, for the volume of literature on the topic is almost literally mountainous) of the research on cloze, is how I attempted to employ the variants of introspective verbal report known as 'concurrent think-aloud' (TA) and, later and due to perceived shortcomings of think aloud, a form of immediate retrospection I have labeled 'non-continuous reporting' (or NCR), to investigate how German and Japanese L1 cloze test-takers/close passage processors approach the task of de-mutilating a cloze passage. The requirements of these two reporting procedures are compared, along with the insights they generated.

Verbal reports as a method of eliciting information about task processing have a fairly long history, and were staple procedures in the analysis of thinking patterns (Freud 1914) and in Gestalt psychology (Duncker 1945) among other fields. Their popularity declined with the rise of an experimentalist paradigm in linguistics, in which techniques and aspirations adopted from 'harder' sciences came increasingly to

predominate, and even to be seen as the only respectable modes of enquiry. In the last twenty years or so, however, verbal reporting has undergone a marked revival, and has been applied to a wide range of language tasks and linguistic contexts. The prime catalyst for this revival has undoubtedly been the 1984 publication (and 1993 republication with revisions) of Ericsson & Simon's '*Protocol Analysis: Verbal Reports as Data*', which presented what many researchers appear to see as a persuasive model of verbal reporting as a methodologically respectable and reliable means of getting at those internal cognitive processes inaccessible by other means.

While few recent studies that have employed think aloud as a means of investigating interaction between informants and tasks fail to cite Ericsson & Simon's monograph in justification of their use of the procedure, my impression is that by no means all of these studies have been faithful to the guidelines set out therein. I am not alone in this: Boren & Ramey 2000 hint strongly at much the same conclusion in their claim that

"[...] descriptions of this practice [i.e., think-aloud] in the [...] literature and the work habits of practitioners do not conform to the theoretical basis most often cited for it..."

In fairness to Ericsson and Simon, it must be pointed out that their 1993 revision to their monograph does appear, to this reader at least, to propose some minor relaxations to their earlier model. As I will try to show in chapters 4 and 5, however, the updated model may still have its problematic aspects. One of these is that (as my experience of using think aloud leads me to believe) the range of authentic verbal report behaviours may be rather wider than Ericsson & Simon allow; as a result, the data elicited does not always fit the fairly tightly-drawn picture they present. (In fairness, again, it is possible that researchers have applied think-aloud procedure to a rather wider range of tasks and contexts than Ericsson and Simon had anticipated.)

A more serious problem attendant on the use of think aloud, however, is the time and effort required for what I have termed the 'pre-analytic' preparation of data. This embraces the actual recording, transcribing and checking of informant protocols

(whose content may be extremely difficult to decipher) before any real analysis of that content can begin. When protocols are in a language other than the researcher's own, of course, the problem is magnified.

Nor, as I shall try to show, is think aloud data inevitably rich and detailed. As Ericsson & Simon's model allows, processing data may be at a level of consciousness below that at which it can be 'attended' – naturally a prerequisite for its being reportable. 'Attention' is a necessary but not always a sufficient condition for reporting, moreover: a pair-condition think aloud informant, for example (I discuss the pair-reporting format in chapter 5) may suppress information which she assumes is already known to her partner, and an informant may fail to report data she herself is unsure of, which she feels will not be interpretable by her researcher-audience, or which may cause her to lose face in some way.

1.1 Alternative procedures

These and other problems with think aloud led me to explore not only the NCR reporting format (see chapter 6) but also the alternative data-elicitation mode discussed in chapters 7 and 8. This is a paper-based procedure I have labelled 'annotated cloze' (AC), in which the informant herself analyses her own processing in terms of an *a priori* set of behaviours derived from the think aloud protocols of previous informants working on an analogous task.

The stages of development of both think aloud and annotated cloze are discussed in some detail, but it remains my hope that this discussion might help a data-collector contemplating the use of either method to avoid some of the errors I myself have made. Some of the benefits, shortcomings and limitations of annotated cloze are presented in chapters 7 and 8 in the belief that, perhaps in some amended or improved form, the procedure may be of value in certain research contexts.

I offer the caution here that the final picture is nothing like as clear as I had anticipated it would be. Annotated cloze may be preferable in some situations and for some purposes, while think aloud or NCR are better suited to others: the appropriateness of each procedure may be only partially predictable, moreover. It is inevitable that some of the small-scale studies and findings outlined here will appear (to put it kindly) less significant to the reader than they seemed to me at the time to me, the investigator. Apart from questions addressed and conclusions drawn, there is a narrative dimension to what I report here, and one of my main criteria for choosing what to include has been simply this: Would it have benefited me to have been told of this before I found it out for myself? To the extent that they provide ..... into how light can be cast ..... test-takers' processing of cloze, the following ..... may be worth a glance.

CHAPTER 2: THE DEVELOPMENT OF CLOZE AS A PROCEDURE

2.0 Introduction & definition(s)

In this chapter I critically discuss the history of the development of cloze, its theoretical underpinnings, and some of its offshoots. We may as well begin with a look at some of the ways cloze has been conceptualized. In 1953 Taylor defined the basis for cloze (which a that time was still perceived as an index of readability) as

> "[the] notion that humans tend perceive a familiar pattern as a whole even when parts of it are missing, obscured or distorted."

Osgood 1959 offered essentially the same definition. Cloze depends on:

> "the tendency in perceptual decoding for observers to fill in or complete familiar forms which are incomplete or obscure as physical events."

but by 1965 the same writer had simplified this to "the tendency to fill in a gap in a well-structured whole." As I discuss later in this chapter, these definitions rely on the notion of Gestalt, or the view that human beings have an innate tendency to construct or reconstruct a picture as an integrated whole. This probably unassailable but not very satisfying conceptualization of cloze, above, does not appeared to have satisfied cloze users for long, and as we shall see fairly strong positions have been taken up over the years on either side of a number of issues. As Stevenson put it in 1979, "[there is] a great deal of uncertainty as to what, after all, a cloze is measuring.", and for Klein-Barley 1983 "a cloze is a cloze is a question."

2.1 Theories Underlying Cloze

Cloze procedure is based upon three key notions. The first and most abstract of these is *closure*–the term used in Gestalt psychology for the innate human tendency to try to impose patterns on perceptual data. This assumed universal urge to complete partial information is still widely used in non-verbal tests of intelligence, where test-takers may be asked to recover a deleted graphic item from a sequence, on the basis of what has come before and/or what follows. Taylor originally (1953) defined a cloze event or unit as

> "Any [...] attempt to reproduce accurately a part deleted from a 'message'[...] by deciding, from the context that remains, what the missing part should be."

but later used more language-specific terms:

> "[...] humans try to complete a mutilated sentence by filling in those words that make the finished pattern of language symbols fit the apparent meaning"

Gestalt theory tells us that humans try to recover apparently missing items, but has little to say about why this should be so. More relevant here is the second key concept underlying cloze as we know it today, the information-theory (Shannon & Weaver 1949) notion of redundancy. Messages typically carry less than the maximum possible amount of information. That is, essentially the same information-content could normally be conveyed via a shorter message utilising the same symbol-set. The unnecessary data is known as 'redundancy', and it is held to be central to successful real-world communication in that in "noisy" environments—in which the message may in some way or other be mutilated in the course of transmission (semi-legible handwriting, intermittent telephone or

data-transfer connections etc.)—the information surplus carried by redundant elements will typically allow the receiver to make sense of the message.

Cloze theory assumes that the more readable a piece of writing, the better it will be understood even if a proportion of the words are removed. Moreover, the better the writing has been understood, the more likely it is that the reader will be able to 'guess' which words are missing. Cloze, then, counts "successful acts of reproduction" (Taylor 1957a:20) and Taylor hoped to have found in it a readability measure that, unlike conventional formulae, took account of meaning. Cloze, then, began as an operationalization of the notion of reduced redundancy, and was intended to measure the comprehensibility of actual texts.

2.2 Cloze as a person-measure

Although cloze originated as a text readability index, by 1957 Taylor and others (Taylor 1957b) had begun to explore its use as a measure of reading proficiency . The US Air Forces Armed Forces Qualification Test was by that point utilising a cloze procedure as a measure of 'mental ability' and certainly by the mid-1960s researchers (Gallant 1965; Bormuth 1967,) were applying cloze also to the measurement of native-speakers' reading proficiency. Before long cloze had come into fairly widespread use as a measure of overall linguistic proficiency in both native and non-native speakers.

Cloze tests then may be used in two main ways. When the test-taker is seen as the constant and text as the variable, we can arrive at a measure of readability. When the text is the constant and the test-taker the variable, we are able to place

test-takers on a scale of—what? Some writers hold that cloze measures proficiency in "higher-order" and "lower order" skills, while others argue that it measures only "lower order" proficiency. Others again say it involves both sets of skills, and yet others claim that it measures neither adequately. The higher-order skills vs. lower-order skills alone debate is mirrored at a less abstract level in the question of whether cloze is to be seen as a test of overall language proficiency, or as a measure of reading skill (and at a less abstract level still in whether or not cloze taps inter-sentential constraints.) I look at these debates in more depth below.

2.3 Cloze theory and theory of language

The Gestalt notion that the "figure" is more than a sum of parts sits well with a unitary view of language. One of the most enthusiastic proponents (for a time at least) of such a view was John Oller's Unitary Competence Hypothesis. Perhaps inspired by the Choimskyan notion of a Universal Grammar, by 1976 (Oller 1976b) Oller was arguing that language competence was an essentially indivisible attribute, rather than an assemblage of component abilities. For Oller, the notion of "pragmatic expectancy" was central to a theory of language. In this model, the central criterion of language proficiency is the extent to which a language-user is able to predict which language items he or she will encounter next in the flow of information. This expectancy is not just a matter of knowledge of language rules (although some, perhaps unconscious, awareness of 'probabilities' must be involved here) but also of the wider pragmatic context. Expectancy grammar, then, is seen as:

> "continually formulating. modifying, and reformulating hypotheses about the underlying structure and meaning of input signals" (Oller & Perkins 1978:44)

Although the label 'expectancy grammar' is Oller's own, the idea seems to have its roots in Spolsky, Bengt, Sako and Atterburns' 1968 observation that dictation tests with added noise (where the noise has the function of mutilating the text) were able to discriminate well between native and non-native speakers. Spolsky et al (op.cit.) explain this by suggesting that what distinguishes non-native from native speakers is precisely the former group's "inability to function with reduced redundancy" and suggests that:

> "[the] key thing missing is the richness of knowledge of probabilities—on all levels, phonological, grammatical, lexical and semantic—in the language."

This idea has a clear implications for attempts to measure an individual's linguistic ability: knowing' a language in any real-world sense must include the ability to deal with incomplete or distorted messages.

2.4 Cloze and test theory

Spolsky 1981 divides the more recent history of language testing into two phases: 1. a 'structural-psychometric' or 'modern' phase, in which batteries of sub-tests attempted to measure a range of sub-skills which were held to collectively make up language ability and 2. an 'integrative-sociolinguistic' or 'post modern' phase, in which tests aimed to measure an assumed holistic language ability. In the first of these stages, a typical test battery might have included tests of grammatical knowledge, phonological discrimination, listening comprehension, and of reading and writing abilities. It was thought that in this way the sub-skills of language

could be measured, and the results assembled into a profile of the testee's overall language ability. If this ability was to be seen as essentially indivisible, however, psychometric tests had to be replaced by measures which could more effectively tap the subject's holistic language competence. The best tests would be integrative tests, requiring rounded or multidimensional use of language in a naturalistic or pragmatic fashion. For Oller & Perkins:

> "a valid language test can be characterised as one that activates the expectancy grammar that the learner has internalised" (Oller & Perkins 1978:52)

Plausible integrative-pragmatic test procedures included cloze tests, dictation, and oral interviews. Pragmatic tests, incidentally,

> "[...]require time constrained processing of the meanings encoded in discourse." (ibid.)

Here, the apparent content validity of cloze and oral interviews was a persuasive factor. The UCH was abandoned, even by Oller himself, only a few years after it came to prominence: a theory originating in factor analysis died by subsequent and more sophisticated factor analysis. The demand for integrative testing, however, was not dependent on the UCH model. Even if language really did consist of component abilities, it could still be argued that these were better tested integratively, in 'authentic' activities such as interviews or conversations, than via psychometric tests involving such discrete point measures as multiple-choice items.

I have discussed elsewhere (Gibson 1993) some of the factors behind the swift rise to popularity of cloze among language professionals. It is enough here to note the attraction of an easy-to-use and, for chalk-face teachers, empowering approach to testing; an approach, moreover, underpinned by the plausible and (unlike factor analyses) readily comprehensible notions of overall linguistic competence and expectancy grammar. Although the notion of 'face validity' has fallen from favour it may be added that cloze seems to have found quite ready acceptance among students (Porsché, pers. comm.) and this apparently on the grounds that it gave the impression either of testing more aspects of language, or of testing these in a somehow more authentic way.

The cloze procedure as we know it today stems from W.L.Taylor's early 1950s work on measuring text readability. The roots of cloze, however, can be traced back at least as far as Hermann Ebbinghaus' 1897 paper "On a New Method of Testing Mental Abilities and its Application to School Pupils". There Ebbinghaus outlined his *Kombinationsmethode*, a procedure which deleted parts of words, entire words or even phrases from a text and replaced them with dotted lines of approximately equal length to that of the original deletion. (That there is very little new under the sun is suggested by the fact that C-Test, 'natural', and 'discourse' cloze focus on the same elements, respectively.) That Ebbinghaus was aiming at what we would now label an integrative or holistic approach to testing—admittedly of intelligence rather than of language ability *per se*—can is suggested by his specification that a test should measure subjects' ability to

combine fragments of text into a meaningful whole. The subjects' ability to do this would be reflected in an ability to fill the original deletions with at least the semantic equivalents of the deletions. (Again, a reasonable definition of one of the two main criteria of cloze success today.)

Harris (1985) notes that Ebbinghaus' method was quickly taken up by intelligence testers in the US. and in Britain. One such, Simpson (1912; cited in Harris op.cit.), found that Ebbinghaus-type completion tasks were among the most effective of the battery of measures he employed in separating subjects according to 'general intelligence'. Simpson's variant of the Ebbinghaus task appears to have been remarkably similar to current rational-deletion cloze, albeit with a rather higher deletion-ratio. Harris (ibid.) reports the administration of a set of Simpson's original (1907) completion tasks to groups of native and non-native speakers of English. The results paralleled those commonly found in cloze testing in that while native speakers clearly out-performed non-natives, the best non-native test-takers achieved similar scores to the lower native speakers. Moreover, non-native scores spanned a wide band, while native-speakers tended to 'bunch' at the top end of the range. Harris goes on to discuss the recasting of Ebbinghaus' procedure as a test of language ability, and notes that these early techniques have "much in common" (ibid.:373) with the cloze procedure outlined by Taylor. Taylor's work may justifiably be seen as an adaptation of an older procedure, and his anchoring of 'cloze' within modern information theory very likely made it more attractive to the language testing culture of the time.

2.5 Criteria for cloze deletion

Jonz 1990:61 describes the construction of a cloze test as

> "..a procedure so uncomplicated that it can be accomplished easily by anyone with a photocopier and a bottle of correction fluid."

The many attempts at 'fixing' cloze procedure, including some by Jonz himself (ibid.) seem to give the lie to this remark. In this section I review some criticisms of cloze which have led to the development of a number of variant task formats, and discuss some of the decisions that may have to be made before an actual cloze task is constructed. These cover the *type* of deletion (true random, pseudo-random and rational) to be used, as well as the *level* (sub-word, word or supra-word) at which deletions are to be made. I go on to look at the 'macro-question' of how a cloze passage text can best be presented, leaving the micro-questions of choice of passage and of deletion ratio to a later section.

Three types of deletion procedure have been proposed for cloze passages: true-random, pseudo-random and rational deletion. The first of these, the *true random* procedure, appears to have been quite quickly abandoned on two main grounds. The first drawback of true random deletion is that it is very awkward to administer. Random numbers are generated by means of a table or calculator, and are matched to words (which have been numbered in sequence) in the target passage. The application of a true random deletion procedure is surprisingly time-consuming, as the set of random numbers generated must be edited to fit the limits of the passage word-count. Finally, the words whose sequence numbers

match the edited random numbers are deleted. The resulting cloze passage may look something like this:

> The study of population statistics is called 'demography'. All advanced _____ now collect detailed statistics of births, deaths and _____. Moreover, every few years a census of population is taken. From a careful study of these figures, demographers have worked out a description of what might have happened in the history of the population of _____ modern industrial nation. Throughout most of human _____, they believe, _____ has had a very high death rate and a very high birth rate.

The selection of candidates for deletion is entirely unpredictable, and it is quite possible that close or even adjacent words will end up being deleted while considerable stretches of text remain unmutilated. The implications of this for the recoverability of deletions are themselves highly unpredictable. Oller & Conrad (1971:163) suggest that true random deletion is more likely than pseudo-random to "bias the test in favour of certain grammatical categories." On the face of it, this suggestion–for which Oller & Conrad offer neither evidence nor rationale–appears to run counter to a basic assumption of cloze theory.

The second reason for abandoning true random deletion was that the more straightforward *pseudo-random* (or '*fixed-ratio*') procedure was held to be just as valid a method. Here the test constructor simply selects a staring point and from there deletes every *n*th word, typically every fifth to seventh word, with hyphenated words generally, but not always, counted as one item. One rationale behind fixed-ratio cloze, then, is that, over a sufficiently long stretch of text, the "random" nature of fixed-ratio deletion will—just like the true random deletion

procedure, if we leave aside Oller & Conrad's claim to the contrary—access all word classes more or less equally. The resulting blanks will thus contain a *representative* sample of prepositions, verbs, nouns, cohesion devices etc.

> "[...].if enough words are struck out at random, the blanks will come to represent proportionately all kinds of words to the extent that they occur." (Taylor 1953:419)

Lexical features ('content words') and grammatical features ('function words') should thus also be accessed proportionately. This comprehensive access to the full range of text-features is, according to cloze theory, what allows the procedure to function as simultaneously a test of test-takers' grasp of grammatical structure and a measure of their awareness of textual cohesion or rhetorical organisation.

2.6 Rational cloze

Although the alternative procedure of rational deletion (i.e. deletion according to some criterion) has been around since the very introduction of cloze as a measure of readability-cum-reading comprehension, it has become much more important in recent years in response to some of the criticisms levelled at the traditional fixed-ratio deletion cloze. (This format has been labeled 'natural cloze' by Brown 1993, and merits a section to itself. Alderson 2000:208 argues that it is the only variant to which the label 'cloze' should be applied, and uses the term 'gap-filling procedure" for what the rest of the world seems determined to call 'rational cloze'.

As the name suggests, rational deletion involves the deletion of items from a text according to some rationale or criterion. Taylor 1957b remarks that almost

immediately after publication of his original (1953) paper commentators began calling for rational deletion on the grounds that this would produce a test of linguistic proficiency which was at once more flexible and capable of being applied with greater precision.

Rational deletion procedure was for many years submerged by what I will call (in a phrase borrowed from Klein-Braley 1981a) the 'a cloze is a cloze' school of thought, well exemplified by the quotation from Jonz, above. In this view, the standard fixed-ratio cloze format represented a perfectly valid test method in need of no improvement or tinkering. Recently, however, rational deletion has been taken up by, among others, Bachman (1982, 1990) and Alderson 2000. It is worth quoting the latter's rationale for his suggestion that the procedure should replace (natural) cloze on the related grounds of efficiency and tester-control:

> "What an individual cloze test measures will depend on which individual words are deleted. Since the test constructor has no control over this once the starting point has been chosen, it is not possible to predict with confidence what such a test will measure: the hope is that, by deleting enough words, the text will be sampled accurately [...] Many cloze items [...] are not constrained by long-range discourse, but by the immediately adjacent sentence constituents [...] Such items will not measure sensitivity to discourse beyond the sentence or even the phrase. Since the test constructor has no control over which words are deleted, she has minimal control over what is tested."

The accuracy of this claim here (echoed in Bachman 1982) will be clear to anyone with sufficient experience of using natural cloze in testing or even in didactic contexts to have perceived the overall pattern of largely local deletion, although

Jonz & Oller (in their 'critical appraisal' of cloze research of Oller & Jonz 1994) seem to downplay the problem of limited discourse constraint.

A perhaps less plausible argument in favour of a rational deletion criterion is that fixed-ratio deletion may fail to access text-features in a suitably representative fashion. It should be noted that this appears to counter Taylor's assumptions about cloze procedure, cited above. Oller & Conrad (1971: 187) have argued that:

> "All that the person preparing the test need do is (1) elect a passage of prose of suitable difficulty definable in terms of a population and its language objectives—of approximately 250 to 500 words in length. (2) Delete every *n*th word (where *n* is usually a number between 5 and 10). This mechanical method of selecting blanks[...] can, in the long run, be expected to reflect the frequency of occurrence of grammatical and lexical forms in the language tested."

Cohen 1980a also suggests that, as fixed-ratio deletion procedures are not equivalent to true random deletion, there is a risk that a word class will be over or under-represented. On the face of it, this appears to be the inverse of Oller & Conrad's claim, above. Again, it is hard to see how a true random procedure could be any more or less likely to bias deletion in terms of the word-class of items selected, or indeed in any other regard.

What *is* likely to bias deletion in favour of a particular word-class is the deletion ratio. As the majority of the individual words in a text will be 'structure' (also labeled 'grammatical' or 'functional') items (cf. Klein-Braley 1985b) a deletion ratio of 1:5 is likely to delete more structure words than is a ratio of, say, 1:7. This has implications for the recoverability of items deleted, for it has repeatedly been

shown (Aborn, Rubinstein & Sterling 1959; Tuinman & Gray 1972) that, by and large (and I will demonstrate exceptions to this rough rule in chapters 6 and 7) structure words are more predictable and hence more readily recoverable than content (lexical) words.

Oller & Conrad's qualification "in the long run", above, echoes Taylor's criterion of "enough words" being deleted. If the total number of words in a text is seen as the 'population', then clearly the more extensive the 'sample' of words deleted, the more likely it is to tap the full range of items the text contains. There is still debate about how many deletions a cloze task must contain in order to be valid, but it has been argued (Sciarone & Schoorl 1979) that a minimum of around 75 deletions are required for the test to be a valid one. Alderson 2000:208 claims the research shows that 50 items are required for a valid test, a figure echoed by Raatz & Klein-Braley 1997, while Rand 1978 claims that only 25 items are needed. The number of blanks a valid cloze test requires may depend in part on the word-items it samples and their attendant constraints, so that intuitively the 50 item figure may be a viable compromise.

Similarly, individual texts used as the raw material of cloze tests must be seen as samples of the population of all possible target language texts. The abovementioned riders of "in the long run" and "enough words", may imply an inkling on these authors' part that the deletions of an individual cloze passage may not in fact adequately sample the target language's lexical resources.

Then again, Oller (1979b:364) claims that

> "In spite of the feeling that just any old text will not do[...]research has shown that the cloze procedure is probably appropriate to just about any text."

This suggestion has been disputed by Alderson 1979; 1981 and Klein-Braley 1981a, both of whom have gone well beyond the realm of 'feeling' in showing that Oller's claim is problematic.

As Alderson 1979a:26 has stressed, a rational cloze procedure must have a rationale on which to base the decision as to which items to delete. Natural cloze relies (Raatz & Klein-Braley 1981) on the notion that pseudo-random deletion will target textual redundancy, and as this is an inherent quality of text *per se* then deletion according to any criterion or none will inevitably target it. Rational cloze/gap-filling procedure, however, typically has the objective of measuring something more than the ability to utilise redundancy, and it is a key assumption of this format that it is possible to select items for deletion in such a way as to target specific aspects of the test-taker's linguistic ability. Fairly common traits to be targeted are (1) the ability to recover 'functional' deletions, thus indicating knowledge of language structure; (2) the ability to recover lexical or content words, indicating overall comprehension of (perhaps topic-specific) passage content, and (3) the ability to recover items which show understanding of local passage cohesion or discourse/global-level coherence.

2.7 Two functional classes of rational deletion cloze

What I am going to call—for want of a better label—*word-class* cloze tests employ the grammatical criterion of word class to delete some or all prepositions, pronouns, copular verbs, conjunctives etc. from a passage. The crude criterion of the deletion candidate's word class membership should not obscure the range or quantity of knowledge potentially required in filling such deletions: while recovering prepositions typically requires linguistic knowledge at the 'local' level of syntax, recovering pronouns or conjunctives may call for rather more in the way of 'global' textual comprehension. As pronouns and conjunctives are important carriers or markers of cohesion (Halliday & Hasan 1976:8ff) it may be difficult to draw a clear line between what I have termed word-class cloze and the notion 'cohesion' cloze discussed below.

'*Cohesion Cloze*' as described by Bensoussan 1990 is distinguished by the fact that test-takers are asked to fill deletions solely from words which have already occurred in the passage. As a result, says, Bensoussan, the Cohesion Cloze "is more constrained and requires students to use text redundancy more closely." (op.cit.:26) Cohesion Cloze, then, makes exclusive use of the passage's redundancy and cohesive relations in order to "evaluate macro-level reading comprehension" (op.cit.) Bensoussan's procedure is not, however, the only procedure which aims to focus test-takers attention onto 'higher', beyond-the-sentence features. As suggested above, 'word class' deletions which remove some or all pronouns, conjunctives etc. may have the same effect, if not the same aim.

Impossible recoveries and targeted testing

Beyond its allegedly more global focus, a feature of rational-deletion procedure is that it allows the test constructor to avoid deleting words which may be effectively impossible to recover, or which require extremely specialized knowledge. This situation is by no means uncommon with fixed-ratio deletion, and in such instances the test constructor is forced to choose between 'bending' the deletion rule (cf. Trollope 1995) or rewriting the passage to displace or replace the offending item (Klein-Braley 1981a; Raatz & Klein-Braley 1981.) The ethics of rewriting texts to ease or enable closure are debatable, but the practice does not seem to be uncommon at least in the didactic use of cloze. In something between an informal sampling procedure and petty theft, I compared ten natural cloze passages prepared by teaching colleagues (and by xeroxing originals left around the photocopier) with identifiable source passages. I found that in only three instances was the unmutilated content presented entirely unchanged, and that in only two of the remainding seven were emendations confined to cutting out of content. In other words, fully half of the cloze passages used had to some extent been rewritten.

A further advantage of rational-deletion is that it allows the test constructor to target responses to a didactically-relevant subset of text-features. These may involve recently taught lexis, likely L1—L2 confusions, or other target items. It has certainly been my experience that such a targeted test has high 'acceptance-value' among students, but the question remains of whether a passage

whose deletions focus on, say, incidental occurrences of last week's vocabulary is much more than a test of memory and/or diligence. Although it has been labelled as such (cf. Cohen 1980a) I would question whether this targeted-deletion format has much to do with cloze as it is commonly understood, and it is not one I intend to discuss further.

2.8 Rational-deletion & cloze as an 'integrative' test format

While one of the major planks of cloze procedure is that it produces an integrative or holistic test task (i.e. one requiring an authentic, multiple-process response from test-takers) it can be argued that some rational-deletion criteria push cloze rather towards the discrete-point end of the continuum: there is little obvious difference between an unashamedly discrete-point test item such as:

> The boy put the box _____ the shelf.
>
> (a) at (b) in (c) on (d) to

and cloze blanks like

> '[..] so Peter went back _____ and put Grandad's mouth organ back _____ its box.'

The adoption of a consciously non-random deletion procedure might, appear to undermine the putative singularity of 'cloze as cloze', and indeed to run counter to the rationale behind it. It may, then, be worthwhile to review more closely the arguments presented in favour of rational deletion cloze/gap-filling procedure.

Bachman's 1982 arguments for rational deletion focus on the findings of Alderson 1979b; Klein-Braley 1981a and others that 'standard' fixed-ratio cloze largely

measures textual 'connectedness' at or around the clause level, so that it chiefly reflects 'lower-level' comprehension on the part of test-takers. Bachman suggests that as more words within a text function at a syntactic/cohesive level than at the level of global cohesion i.e. coherence, random deletion will inevitably sample a larger proportion of "clause-bound" words. (The assumption here seems to be that we would not specifically *wish* to target clause-level syntax or cohesion.) The inconsistent results identified by Alderson 1978, Klein-Braley 1981a and others, then, can to some extent be explained by this unbalanced focus of standard cloze:

> "One possible reason for the inconsistent results of previous research may be the adherence to the principle of random deletion [...] Random deletion ignores the syntactic and semantic relationships in a text, and is therefore likely to yield inconsistent results." (Bachman 1982:66)

Cloze passages mutilated according to a rational deletion procedure can, according to Bachman, be used to measure textual relationships beyond the level of individual clauses. Clearly, Bachman is here moving away from the earlier view of cloze (Bormuth 1967 Oller 1973, 1975; Ramanauskas 1972) as a test of linguistic proficiency. Bachman suggests that:

> "The advantages thus gained may well offset considerations for measuring random redundancy or general proficiency." (op.cit.:66)

The implication behind this suggestion appears to be that cloze as a test of reading comprehension and/or readability and cloze as a test of overall language proficiency may be quite different creatures.

2.9 Rational deletion cloze, C-Test & general language proficiency

But rational-deletion cloze tasks have also been seen as valid indices of overall language proficiency. Hughes (1989:66) reprints a rational-deletion cloze test similar to those which have been used in the Cambridge Proficiency Examination. The deletion-criteria for this test were, Hughes reports, twofold: on the one hand, deletions were chosen to provide 'interesting' items (although what this might mean in practice we are not told) and on the other hand to test students' ability to process "various features of context" beyond the level of grammar. Hughes notes that when used as part of the CPE test 'battery' this type of cloze task produced scores which "correlated very highly" (again, no actual coefficient is given) with the CPE as a whole. Rational-deletion tests are, Hughes argues, the cloze format to be recommended as a measure of overall language ability.

If we include the C-test format (Klein-Braley 1984b; Klein-Braley & Grotjahn 1995) under the umbrella of 'clozoid' testing (and these originators of C-Test explicitly see it as measuring the same ability as natural cloze, i.e. applying linguistic redundancy) then the 'levels' at which items are deleted range from the C-Test's deletion of the second half of every second word (which prod____ something li__ this) through 'standard' fixed-ratio every–*n*th– word deletion to 'discourse cloze' formats in which the items targeted for removal are phrases or other textual units. Word-level deletion has already been taken up, but as both C-test and 'discourse' cloze were developed in response to perceived shortcomings of conventional cloze it is worth looking in some depth at these variant procedures.

The C-test format is discussed in detail in Raatz & Klein-Braley (1981, 1983), Klein-Braley & Raatz (1985) and elsewhere. This test was developed at the University of Duisburg in response to a number of perceived shortcomings of standard cloze (Klein-Braley & Raatz 1985.) Like cloze, the C-test is intended to be a test of overall language proficiency operating via reduced redundancy, but with the following features either lacking or unreliably present in cloze:

1. Several (ca. 5-6) different passages are used
2. At least 100 deletions are included, which [sic] "affect a representative sample of the text"
3. Native speakers should achieve "virtually perfect" score
4. There should be no need for any other scoring method than 'exact word'
5. The reliability and validity of the test should be high.

As in cloze test construction, a 'lead-in' is left unmutilated: in the case of C-tests the first sentence should have no deletions. From this point on, the second half of every second word is deleted until the target number of deletions is reached. From there on the text continues without deletions to a "natural break" (op.cit:136) such as a paragraph boundary. Klein-Braley & Raatz further claim that the C-test appears to do "everything that the cloze test promised", so that it may seem surprising that cloze has not been entirely superseded by the C-test format. The alleged advantages of C-test over standard cloze are numerous, and are given here in no particular order:

1. The larger number of test items per x words of text boosts efficiency and reduces the risk of bias through subject-fatigue.
2. The presentation of several texts reduces the risk of bias from text content.
3. Very high reliability and validity coefficients are consistently obtained.
4. Test construction and scoring are extremely easy. In addition, scoring is highly objective.
5. The test has a high face-value.
6. Adult native speakers typically achieve (near) perfect scores.

Cleary 1988, whose own paper discusses an attempt to overcome the poor discrimination of at least some C-tests, notes the serious lack in the literature of any systematic critique of the format. This is a fair comment insofar as the bulk of work on C-Testing appears to have been carried out by proponents of the procedure in the course of attempts to refine it. Moreover, despite the claims of those behind the development of the C-Test, the procedure has not become widely popular outside Germany with anything like the speed its alleged advantages would lead one to expect. Unlike cloze—essentially an English-medium development—the language of development and propagation of C-Testing was at least initially German; this may be a factor in its slower diffusion. As with cloze , the bulk of the research into C-Testing has been devoted to quantitative, statistically-based attempts to validate the format as a test of overall language proficiency. However, on the assumption that the cognitive processes stimulated by cloze and C-Test-taking may have something in common, I look here at some research into these.

C-Test and cloze

It is plausible to expect that the overlap between cloze and C-Test-taking processes will be quite marked. Both, after all, claim to be measures of overall linguistic proficiency, operationalising redundancy via textual mutilation. Both, moreover, provide cues via context and co-text. A key difference lies in the C-Test's inclusion of the initial halves of every target. This (Feldmann & Stemmer 1987) is widely seen as an important source of cues. In the abstract, then, it is tempting to see cloze and C-Test as broadly similar in the cues they potentially offer the test-taker, except that C-Test offers an additional contextual-morphological set of cues unavailable in cloze test-taking. But what do we know about real-world cue-uptake by test-takers?

Although Klein-Braley (pers.comm.) has cast doubt on the usefulness of verbal reports in getting at C-Test-taking processes (preferring to use an approach I discuss below), Feldmann & Stemmer 1987 have made use of verbal data. These writers echo in regard to C-Test the comment of many critics of cloze testing, that we still have no real idea of what the procedure is in fact measuring. To put it another way, we need a clear idea of its construct before we can assess its construct validity. Feldmann & Stemmer look first at the kinds of 'retrieval cue' available in C-Test processing. They accept that the most obvious cues are the initial halves of the texts' incomplete items, but argue that co-text and context will also provide important aids to retrieval. Feldmann & Stemmer (op.cit.:255) accept, too, that the degree of redundancy obtaining will vary among deleted elements.

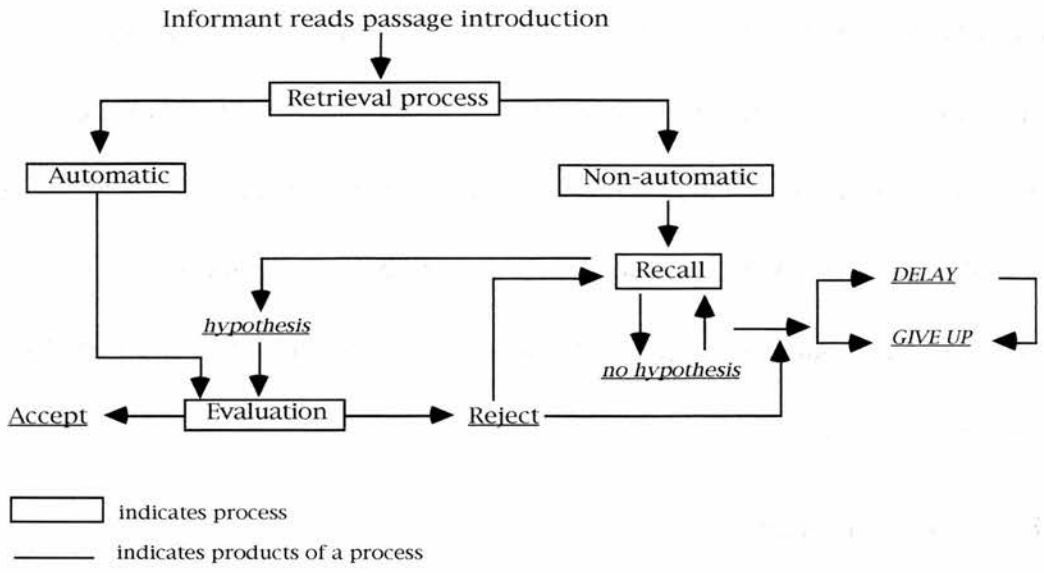

Figure 2.1: 'A tentative model for the problem-solving process in C-Test taking' (after Feldmann & Stemmer 1987:257)

'Automatic' [sic] retrieval occurs very quickly and "without 'thinking'." Instances in which subjects showed no sign of reflecting on an item, and required less than 2 seconds to recover it were classed as instances of automatic recovery. No reasoned basis is presented for selecting 2 seconds as a time-criterion (although my own experience suggests that under normal conditions of reporting this interval is close to the minimum practically measurable) and it could be argues that a more reliable criterion of 'automatic' processing might be that of an informant's inability to recall—in more or less immediate retrospection, prompted perhaps by the researcher—his or her problem-solving behaviour. A given deletion may, after all, be quite easily recovered in less than two seconds. Alternatives to the fairly high-inference term 'automatic' might be 'high speed-and-unattended', 'high speed-and-unreported', or even 'high speed-and-unreport*able*.' These minor points aisde, Feldmann & Stemmer's model,

though tentative intuitively seems to describe many stages in cloze-processing fairly well, although the process labelled 'recall' should not (in my view) be interpreted as implying a more-or-less passive act of remembering. In a real sense, 'recall' of linguistic or 'world' knowledge is *the* necessary condition of cloze or C-Test success, but—as I shall try to show—this is often an active and partially traceable process.

2.10 Discourse cloze

Cloze-like tasks in which deletions are made at a supra-word level have come to be known as 'discourse cloze' tasks. Bensoussan 1990, Butler 1991 and others have constructed rational cloze tasks in which deletions are restricted to 'cohesion markers'. Following—at least implicitly—the argument that standard cloze very largely measures what Alderson 1979 labels "lower order" comprehension skills (i.e. limited to the level of clause or sentence) this variant is intended to force the test taker to recover items cued at 'discourse level.' Widdowson 1989 makes the distinction between text and discourse that the former is a product of "the language system", while the latter pertains to "communication as a whole". Deyes 1984 bases on this his argument that deletions drawn from the set of 'textually cohesive' items (cf. Halliday & Hasan 1976) create a test task which function at a *textual* rather than at a true discourse level. Deyes goes on to argue that by drawing candidates for deletion from a closed set of cohesive items, we lose the option of scoring as correct semantically-acceptable, rather than solely exact-word recoveries. In Deyes' view the closed set of cohesion markers is made up of items which are "mutually exclusive, and where distinctions are often subtle"

(op.cit.:129) he cites an example from a 'discourse cloze' produced by other writers (Levenston *et al*, cited in Deyes, op.cit.):

> "The Queen said 'Curtsey while you're thinking what to say. It saves time.' Alice wondered a little at _____, but she was too much in awe of the Queen to disbelieve it."

and their comment that only 'this' shows an understanding of how an immediately–preceding item of text is referred to in narrative prose.

What Deyes is saying here is clearly true in part, but it seems excessive to suggest (if I have understood his meaning correctly) that semantic distinctions among cohesive items are invariably subtle. Nor can we ignore the element of user-choice here; I would argue that in a context such as the one above there is in modern English little or no restriction on choice of 'this' or 'that', despite a residual preference for the former in more formal register. This seems to undermine the idea that cohesive items are mutually exclusive and scorable only via the exact word criterion, as would the argument that *conjunctive* cohesion markers too (cf. Halliday & Hasan 1976:227ff) can be interchangeable, as in

> She found his chequebook despite/in spite of the mess the room was in.

But to return to the main point of Deyes' critique of Levenston *et al*—that their 'discourse' cloze is in reality a 'textual cloze'—we find that Deyes' solution is to require test-takers to replace not single words but 'communicative units'. Deyes argues that by having test takers recover "'information' rather than linguistic items" (op.cit:130) the resulting cloze task is at a true discourse level. The gist of

Deyes' suggestion seems to be that supra-word strings should be deleted on the criterion of information-content, but such a standard would not require deletion of a string of words where a single word itself carried a reasonably high information load:

> "Had anyone taken the time to ask her, Mariko would have admitted that she felt nothing but ..... at the prospect of marriage. Her parents took a rather different view."

It is not obvious from the extract above (taken from an unpublished translation of a Japanese short story) that the missing word is 'happiness', and this would not become obvious until one had read three full (and fairly long) paragraphs further. Intervals between deletions, then, may be as important as (to coin a phrase, perhaps) the content of the blank.

A further attempt to construct discourse level cloze can be found in Butler 1991. Butler defends a variable deletion ratio on the grounds that the test designer needs to exercise skill in "anticipating what degree of subtlety of meaning the candidate can be expected to cope with at [a] particular level." Butler clearly would not agree that a valid test can be constructed simply by applying straightforward mechanical deletion to more or less any text. The solution proposed is to create, with the help of concordancing software, cloze-like tests in which each item consists of four sentences, from each of which the same word has been deleted. The filler selected must be acceptable in all four contexts:

> (a) Such an approach is usually the ..... of choice for buying the best car.
> (b) I had to live with this ..... for nearly two years.

(c) This is not the ideal ..... for a student to check his or her progress.

(d) This is a common ....., even though many people fail to appreciate that such analysis represents an integral part of the process.

As each sentence has been taken from a distinctly different text, quite how this leads to a 'discourse' cloze is not immediately apparent. Butler (op.cit.) argues that

> "The [sentences] are not isolated, simply taken from a large body of different texts in which the deleted item functions as an integral part."

'Discourse' seems here to refer to the 'textual universe'—an interpretation rather at odds with the usual conception of the term. Butler reports no trial of his proposed format, so we have no real idea of how it has been received by learners. Acceptance of the format might, however, be higher if it were used as a classroom means of looking at the subtleties of word-meaning, rather than as a test method.

2.11 Strong and weak positions

The necessarily brief overview of some of the main offshoots from 'basic' cloze procedure leaves many questions unanswered. My main intent here, however, has been to show that cloze procedure is not the perfected technique that one might assume from Jonz' quote at the beginning of the chapter. Virtually since its introduction in a modern form by Taylor 1953, cloze has been questioned and often found lacking, and much effort has gone into remedying its perceived shortcomings.

Some commentators have noted and decried the at times highly adversarial tone of the debate (Klein-Braley 1981a) over whether or how well cloze-as-cloze targets the full range of textual constraint, but given what is at stake this intensity is not altogether surprising. The 'strong cloze' position in Oller & Jonz 1994, for example, is that (with some under-specified limitations) any text and mechanical-deletion procedure will produce a valid test. Opposing this view are a number of studies including Alderson 1978; Smith-Burke *et al* 1978; Klein-Braley 1981a; Bachman 1982 and Shanahan, Kamil & Tobin 1982. Opponents of the strong cloze position appear at times to (over)state their case without due care. Bachman 1983:66, for example, makes the claim of cloze that:

> "[…] a random deletion procedure would tend to sample a larger proportion of clause-bound words, and]so] appear to be measuring *only* lower-level skills."

The italics are mine, and reflect my unease at this juxtaposition of ratio and exclusivity. Cloze-skeptics have been known to argue (cf. James 1979) that studies such as Alderson 1978, Porter 1976, and others have proven that the procedure measures low-level skills and/or uptake of local cues exclusively, and this also seems to overstate the case. Oller & Jonz 1994: 372ff point out the logical error involved in of trying to prove a null hypothesis empirically, but is that what Porter 1976, Alderson 1978, 1979a; Klein-Braley 1981a, and other primary researchers were in fact trying to do? Or was it rather that they were attempting to show that the null hypothesis about cloze–that its 'real world' sensitivity to long-range textual constraint is in fact quite limited—was tenable at least part of the time?

My own interpretation of these authors is that what we might term their 'weak cloze' position is rather that *certain* passages containing *certain* deletions will *to some extent* target longer-range constraints, but that this effect is *unpredictable*. This conservative claim (which appears to be close to that expressed in Alderson 2000:208ff) is in itself quite enough to undermine the strong cloze position, for once the unity and universality of cloze become untenable the weakness of the fallback position, as it were, becomes plain: cloze *sometimes* operates beyond the local level; just as often, if not more often, it does not. That the implications of this conclusion are not lost on the proponents of cloze is, I think, adequately evidenced in the chapter entitled 'Critical Appraisal of Related Cloze Research' in Oller & Jonz 1994. To this reader, that review blends cogent critique with minor, even trivial and carping criticisms, yet appears to treat both as equally powerful ammunition in defence of the strong position. It may be some time before the dust settles on this question.

CHAPTER 3: CONSTRAINT & STRATEGY

3.0 Introduction

In this chapter I look at two concepts which are inextricably linked: the notion of constraint on interpretation of text and thus of choice of cloze recovery, and the strategies or operations through which readers and cloze test-takers interact with and realize that constraint.

3.1 Constraint & where it comes from: some definitions

I turn first to the notion of textual constraint on interpretation, which as it were provides the bridge between the abstract notion of 'closure' and the concrete interaction of subject and text. In the literature on lexical inferencing and on cloze the term 'constraints' occasionally carries a negative connotation, i.e. indicating what a string cannot mean, or what cannot fill a deletion, as opposed to the positive sense of 'cues' or 'clues' to what the string could mean, or to what might fill the deletion. For the sake of simplicity (and also because a distinction between positive and negative cues is not always workable in practice) I will use the term 'constraint' to refer to the abstract delimiting effect of textual or extratextual information on potential meaning, and take the terms 'cue' and 'clue' (used here interchangeably) to refer to specific, often concrete, textual or extratextual items which offer some guide to what (in that context) a deleted item can or cannot mean.

Contextual constraint is seen (Cziko 1978) as a key aspect of reading ability, so that readers who are unable to make use of textual constraint will be unable to effectively exploit the texts with which they are confronted. But what is textual

constraint? Where does it come from, and what makes a reader able or unable to utilise it? The literature on constraint is very extensive indeed, and it spans a wide range of contexts from phrase and sentence contexts up to full 'authentic' (reading) passages. It also encompasses a variety of tasks including lexical decision (in which the subject has to decide whether the string of letters presented does or does not represent a word), various sorts of gap-filling task. and lexical inference tasks in which the subject has to work out the meaning of an unknown word or phrase within a sentence or passage (Haastrup 1991; Aizawa 1998.)

Much of the work on constraint to date had utilised single-word contexts (DOCTOR → NURSE) and highly predictive single-sentence contexts (THE CAT WAS CHASED BY THE → _____.) and there are doubts (Kamil *et al* 2000) as to how far we can apply findings from such tasks to the study of real-world reading. It is intuitively plausible, however, that the means by which a subject fills a gap at the end of an isolated sentence have a good deal to do with how she fills a gap in a continuous text. It may then be rather more legitimate to apply the insights gained into how constraint operates in experimental contexts, such as single sentences, to the study of cloze processing–even if we choose to be more circumspect in applying these to 'authentic' reading contexts.

3.2 Not all constraint or processing is 'conscious'

A number of writers have identified 'preconscious' or 'attentionless' processing as an important aspect of text decoding. Dixon 1981 suggests that:

"...a large part of sensory inflow which has undergone full preconscious processing up to a semantic level will never achieve conscious representation."

Words, says Dixon. may be semantically analysed at a preconscious level, and what individuals subjectively experience as 'guessing' may in fact have been constrained by stimuli of which they were not fully conscious. Greenwald et al 1995, posit the claim that processing of stimuli may be 'detectionless', i.e. this processing *cannot* become part of consciousness. The implication of the above for introspective studies of constraint is obvious: if some processing is carried out below the level of conscious attention, we cannot hope to access it even via the processor's self-reports. Verbal report data will therefore necessarily be incomplete.

Parafoveal information

The area of parafoveal vision lies just outside that of foveal vision, i.e. the eye's zone of selective focus, and is believed to be a source of–perhaps not fully 'conscious'–textual information. Taft 1991 discusses the interaction of parafoveal information and contextual cues as these affect the word recognition (lexical decision) task, but I know of no study that has successfully analysed the impact of parafoveal information on less controlled reading contexts. The realisation that readers apparently take in and process more than they are consciously aware of has been an important factor in the decline of the "psycholinguistic guessing game" model of reading (cf. Goodman 1967; Ellis 1984.)

One concrete implication of the availability of parafoveal information for verbal report data is that it becomes very difficult to know with any real accuracy exactly when a subject encounters an item of textual data, or a deletion. That deletions exert a powerful 'pull' on the eye is suggested by my finding that subjects to whom I allowed 5 seconds exposure to a short cloze passage ('THE BLANKET') were better able to assess the number of unnumbered, underlined blanks it contained than they were able to assess the number of sentence openings.

The only means I have encountered of establishing precisely when a subject has at least potentially visually encountered a text item involves the use of an eye-movement recorder (EMR) in combination with an individually established span of parafoveal vision. Use of EMR equipment, however, requires that the head by fixed securely in position via a bite-bar, and verbal reporting in such a setting is clearly impractical. A 'low-tech' alternatives to EMR, in which informants are asked to trace with a pencil the part of a passage they are attending at each moment was used in Feldmann & Stemmer 1987 and is presented there. My own trial of this procedure, however, suggested that, while some appear to do this 'naturally' most informants do not consistently trace the shifting foci of their attention. Moreover, the procedure clearly imposes an additional burden on subjects, further removing the task from an authentic reader-text interaction.

3.3 Constraint is cumulative

Constraint is cumulative in that the range of possible continuations declines as a clause progresses:

1. The dog
2. The dog chased
3. The dog chased the

While the choice of clause completions is wide open in [1], it is limited in [2] to either a verb post-modifier (happily/frantically) or–perhaps more plausibly–to a noun-phrase denoting an entity capable of movement (the cat/the postman/his tail), and [3] more or less rules out anything but a noun-phrase. While the first context can cue only the most general prior knowledge of dogs, information from relevant schemata (possible dog behaviours) interacts with linguistic information (possible predication) to delimit the range of potential 'next items'. Lutjeharms 1994:150 uses the term 'valency' to describe:

> "the potential of words as units of the lexicon to combine with other units in sentences".

It follows from the above that a word's position in the sentence at least–the effect of passage position has perhaps not yet been adequately explored–is one factor in how contextual constraint operates on it.

Taft 1991 sees two effects of sentence context on word recognition: context (1) facilitates the recognition of a small set of likely 'next items', and at the same time (2) inhibits a larger set of unlikely items. Potential 'upcoming items' are facilitated if they are 'congruent' with the text, while non-congruent items are

inhibited. In the absence of some deeper insight into what exactly 'congruence' is and how it comes to be, however, this claim is perhaps more a high-level description than an analysis. As facilitation–and hence presumably inhibition–is supposedly not an attentional process, it is difficult to see how it can be investigated in realistic contexts.

Perfetti, Goldman & Hogaboam 1979 see contextual constraints as examples of redundancy. In a text, they say (op.cit:273)

> "the effect of any semantic constraint...is to decrease the number of likely alternatives within the text"

This is plausible enough, although the question of whether 'valency' is operationalised in reader-text interactions as 'inhibition' or as 'facilitation' appears still to be open. My own informants thinking-aloud about cloze deletions did on occasion report that the item *could not* be X or Y, but more commonly they verbalised a list of possible fillers. The situation in which a large proportion of informants do report what an item *cannot* be is when subsequent information in the text makes clear than an earlier filler was in fact wrong. In retrospective interviews, however, informants may volunteer the information that they were aware at the time the missing item *could not have been* X or Y, even though they could not think of a suitable filler. It may be impossible to determine whether such comments stem from genuine 'on-task' insights or are later reconstructions, as even the informant herself may not be in a position to know this.

3.4 Not all contexts are equal: passage difficulty & familiarity

Perfetti *et al* 1979 note that contexts can differ widely in the degree of constraint they provide. In Graesser's 1981 discussion of 'passage familiarity', i.e. prior knowledge of the passage's topical content, he notes a lack of evidence that greater familiarity with a passage content leads, as one might expect it to, to a shorter reading time, and suggests instead that greater familiarity with a passage topic may actually encourage a deeper engagement with the passage, resulting in a longer reading time.

Passage difficulty

A reader's difficulty with a passage may stem from its conceptual or 'content' (see above) or from its linguistic or structural complexity. Mogge 1979 offers an excellent discussion of how linguistic complexity can affect the reading of German texts with fairly straightforward content. The finding that linguistically 'easy' passages are not inevitably read more quickly could be taken as suggesting that (as may be true of topic familiarity) relative ease in reading a passage may lead to a deeper involvement with it.

Rhetorical knowledge

Operating alongside topic familiarity and the level of linguistic difficulty may be 'rhetorical knowledge' which allows the reader to identify the 'text type' or passage genre and predict its structure. Wolff 1989 claims that text-type identification is "an important...processing strategy", is commonly found in informant protocols–taking place before, or very early in, text processing, and

relies heavily on non-linguistic prior knowledge. Wolff does not tell us whether informants identified text-types entirely spontaneously in his study, and I am personally sceptical of his claim that text-type identification is the second language reader's "first and foremost" strategy. That said, where a reader already possesses knowledge of a text genre's typical structure(s), this almost certainly facilitates her comprehension to some degree. Certain other regularities have been observed in the way constraint is activated by text processors.

Context preceding the target is more accessible than that following

DiVesta et al 1978 (cited in Klein Braley 1981) compared cloze tasks requiring reference to constraint in the context preceding the deletion only, and also to constraint in the context following the deletion. Less proficient readers were found to be less able to utilise constraint in the latter context. As we shall see in chapters 6, 'searching beyond' a target deletion appears to be a less-common behaviour.

The distance between target and constraint may affect uptake

It is widely accepted (Sasaki 1993) that textual cues which are physically closer to the target item are more readily taken up than are more distant constraints. A tentative examination of verbal report data suggests, however, that the role of this textual distance factor varies markedly among subjects. Any worthwhile taxonomy of textual constraints and their uptake must provide a means of recording the distance between the target and any textual cue(s) activated.

Not all readers gain equally from contextual constraint

Perfetti *et al* 1979 found (using a lexical decision task) that the availability of context, and hence of constraint, benefited comprehenders at all levels of proficiency. Cohen & Aphek 1981, however, found that more proficient readers of text derive greater benefit.

Prior knowledge more constraining than textual data?

Numerous taxonomies or schemes of constraint have been developed, and despite their differences these tend to have certain elements in common. Two potential sources of contextual constraint are more or less universally recognised: (1) constraint contained within the text, or *intratextual* constraint, and (2) that which stems from *extratextual* knowledge. Extratextual knowledge is often subdivided into general knowledge, which includes linguistic knowledge, and specific topic knowledge, although this division has been questioned (Taft 1991) An example of a deletion recoverable via intratextual (in this case syntactic) constraint would be this:

(1) If I were you, I _____ look for an easier topic.

while a deletion requiring extratextual constraint would be:

(2) The great fish thrashed as the _____ entered its flesh.

Researchers, not all of whom (cf. Taft 1991) were working with text level stimuli, have presented diverse interpretations of the relative importance in reading comprehension of constraint from prior knowledge and that from the text itself (cf.

Graesser 1981; van Dijk & Kintsch 1983; Carver 1994.) It seems only logical that the roles of these two sources of constraint must vary across reader-text interactions, however, so that Harris & Monaco's 1978 suggestion (cited in Taft 1991) that 'pragmatic inferences' derived from prior knowledge and 'logical inferences' derived from text statements combine to create a 'running context' of the passage may be as much of a generalisation as can safely be made in talking about this aspect of the interaction of readers with authentic text.

3.5 Cziko 1978: a more elaborate taxonomy of constraint

Cziko 1978 offers a tripartite division of constraint into (1) *syntactic*, constraint, based on rules of grammar (as in the requirement that the only verb form that can follow *I think he is..* will be in the present continuous) and (2) *semantic* constraint, based on the restrictions or expectations imposed by lexico-grammatical features. For example, the role of [+AGENT] as in *The small dog* leads us to expect that a verb phrase will follow, and *The small dog + stole* will, probably, be followed by an [+OBJECT] noun phrase. Category (1) above clearly represents straightforwardly intratextual constraint, but already (2) partakes of 'knowledge of the world' in that the objects open to theft, by dogs at least, are typically physical objects such as sausages.

Cziko's third category, 'discourse' constraints, defined as '[those] provided by the topic of the text' (1978:473) are very much open to 'general' knowledge of the world and potentially also to specialised 'topic' knowledge. Cziko was attempting to show that, while even low-proficiency readers of a second language

can exploit syntactic constraints, a fairly high level of proficiency is required for effective use of the language's discourse constraints. If we accept, however, that the utilisation of discourse constraints in foreign-language reading depends in significant part on extratextual information, we might choose to rephrase this conclusion in terms of knowledge rather than of proficiency. Categorisations like Cziko's might be seen as 'macro-level' divisions. As we shall see, a rather finer taxonomy of constraints is called for in looking at how readers utilise constraints in practice. I take up below first intratextual constraints and then extratextual cues.

3.6 Intratextual Cues

The term 'cotext' is widely used to refer to the totality of the text surrounding the target(s) of attention. Although 'cotext' is generally contrasted with 'context'–those sources of information outside the actual text–the usefulness of this distinction is open to question. It is now agreed (cf. Widdowson 1989) that the comprehension of a written text involves an interaction between the information which is 'physically present' on the page and that which is 'brought' by the reader. If different readers bring different kinds or amounts of information to a physical text, then that text in itself can have no fixed meaning. Thus the cotext has *potential* rather than actual meaning, and this potential meaning will vary unpredictably depending on the linguistic proficiency, affect, and anything else the individual reader brings to the comprehension task.

But to what extent does the end-interpretation derived from this blend of textual and reader-internal information vary in practice? That different readers can arrive at different interpretations of a given text ("Thou shalt not kill", for example) shows that this potential variation exists; that different readers can derive highly similar meanings from a text shows that it is unlikely to be infinite. While there exist various techniques (multiple-choice, true-false, gap-filling etc.) for establishing the degree to which readers' interpretations of a text can vary, perhaps only verbal report has the potential to show us the process by which meaning is constructed. (The think aloud data shown in chapters 5 and 6 includes instances in which individual readers construct different models of meaning based on the same spans of physical text.)

Intratextual cues have been perhaps most extensively investigated in the field of lexical inferencing (Haastrup 1987; 1991) or the construction by the reader of meaning for unknown words in a text. In lexical inferencing, intratextual cues come from two main sources: the morphology of the target item, and the surrounding 'cotext'. While target-item morphology clearly plays a role (Klein-Braley & Raatz 1985) in such cloze offshoots as C-Test, this does not hold in fixed-ratio cloze: here all morphological cues are (provided all deletions are off equal physical size) completely obliterated.

One of the central debates (a "highly adversarial" one, claims Klein-Braley 1981) in cloze testing revolves around the still very much open question of whether

cloze taps cue-uptake, i.e. constraint, at a 'local' or at a 'global' level. That is, whether cloze deletions are recovered on the basis of textual constraint contained within the same clause or sentence as the deleted item, or at a much wider textual level. Too many claims and counter-claims have been made to be discussed in any detail here (see Oller & Jonz 1994 for a fairly exhaustive survey) but would point out–a view which Alderson (pers. comm.; see also L-TEST-L discussion list November 2001) appears to subscribe–that it is hard for test-constructors to be sure which potential textual cues (if any) will in fact be taken up by subjects; it is thus problematic to assume that test-takers 'must' have used a particular cue or set of cues when recovering a deletion (cf. the discussion in chapter 5 of German and Japanese L1 consultants' predictions of cue salience.) To the extant that the cues actually used (or not used) are made apparent, the verbal reports of informants may be of some value in test construction.

3.7 Extratextual Cues

We need a label for that information which exists independently of any given text, and probably the most widely-used terms here are 'background knowledge' and 'prior knowledge'–of which I prefer the latter. Prior knowledge, conventionally seen as structured into 'schemata', 'frames', or 'scripts' (cf. Pressley & Afflerbach 1995) is taken here to refer more to general and specialised knowledge of the world than to knowledge of language *per se*., although it is not clear how far these can be distinguished. Two things are worth noting at this point:

1. Comprehenders will bring different 'bundles' of prior knowledge to a task.
2. Prior knowledge 'possessed' need not match that 'applied'.

It is extremely difficult to predict accurately what prior knowledge a subject can bring to a task, let alone how much of that knowledge she will activate. The subject, after all, may not be aware prior to its activation by the task of the knowledge she has available. We can, however, arrive at a rough and partial assessment of the subject's prior knowledge 'potential' by means of pre-task tests of general knowledge. Such pre-tests have been utilised (Ridley 1997) in an attempt to control for variation in subjects'prior knowledge.

It is plausible to expect that a subject who scores low on (say) a multiple-choice test of knowledge about 'the Olympic games will also experience more difficulty in completing a clozed version of a text on that topic. If cloze–success is influenced by prior knowledge, then we ought to see correlations between such scores on such pre-task tests and on the associated cloze passages. (One difficulty, however, is that by posing a test of general knowledge prior to a data-elicitation cloze task, the researcher risks alerting potential informants to the task topic. If the order of presentation is reversed, the informants will have acquired topic knowledge from the task.) It also becomes clear from think-aloud protocols that prior knowledge may be potentially 'available' (in the sense of being stored somewhere in long-term memory) without being, if you like, sufficiently activated by a given task. Informants are frequently able to produce a piece of information in post-task interviews which they were unable to produce in the task itself.

Remarkably, Conrad 1962 found subjects who appeared not to know *post*-task words which they had known *pre*-task, and suggests that "the vocabulary of the individual is not always the same, or always equally accessible." The individual's topic knowledge and her linguistic knowledge appear to be tapped differently in different contexts and tasks, and Widdowson's (1989) suggestion that written text is essentially a series of pointers to which bits of knowledge to activate implies that some texts 'point' more effectively than others in individual text-reader interactions.

3.8 Topic Knowledge

Beyond the general knowledge of the world that we expect most readers of a given linguistic and/or cultural background to have internalised, there is the specialised knowledge–or rather *knowledges*–possessed by various sub-groups within that background. These knowledges may be of a particular topic (horse-riding, electronics) in which we can expect 'initiates' to have registered the existence and contextual meanings of certain lexical items and phrases. Thus, only a sub-group of even quite proficient L1 readers could be expected to know the term *rheostat*. Knowledge of a word's meaning in one specialised context may be of limited help, moreover (cf. Haastrup 1991) when the word is encountered in another context: a sailor's *sheet* and a carpenter's *sheet* are two quite different things, although neither has to do with paper or beds.

3.9 What Do We Mean By 'Strategies'?

In this section I look at the notion of cognitive 'strategy'. I discuss some of the problems involved in defining and operationalising the construct of strategy, and in categorising the behaviours identified. It is worth pointing out here one immediate problem with the term: for many, 'strategy' carries the connotation of conscious awareness–even of 'deliberateness'. In the research literature, however, this cannot be taken as read.

There is inevitably a temptation to apply the more rigorous and 'commonsense' criterion of 'deliberateness' in labelling a behaviour as 'strategic'. Unfortunately, it is extremely difficult in practice (Taft 1991) to distinguish deliberate from non-deliberate behaviour. Even separating conscious and unconscious behaviour (ibid.) is far from straightforward, so that even this criterion is problematic. Standard usage in the field seems to be to accept that the label 'strategy' covers both deliberate and non-deliberate actions (or indeed unattended and/or unconscious actions) and can even be extended to cover 'non-events' (Smith 1994) such as failure to do *x* or heed *y*. My own prejudice is that the term 'strategy' has come to take on so many disparate meanings and connotations as to be almost unusable without careful definition or delimitation. One cannot buck the *Zeitgeist* entirely, however, and so I retain the term 'strategy' when discussing the work of other researchers. In discussing what I have observed my own informants to do and say during their task-processing, however, I propose to substitute the more neutral terms 'behaviour', 'operation', and 'event.' (These are used for the

most part synonymously.) I will do my best here to avoid applying the term 'strategy' to my own informants' behaviours, but if it does creep in it should be taken to imply conscious and deliberate application of a particular choice of action.

3.10 The investigation of strategies

It is worth looking at the notion of strategies in more detail, for the concept is a distinctly slippery one. First, a quotation from Skehan (1989:98):

> "If...we review the whole of ...strategies research, we have to say that the area is at an embryonic stage. Conflicting results and methodologies proliferate. There are few hard findings. We seem...to be dealing with a...research-then-theory perspective, in which there is no established framework for the research which has been conducted, and in which different investigators 'trawl' in different ways."

It is hard to argue with these claims, although it is only fair to point out that the 'research-then-theory' approach (which in some variants does indeed lack a consensual framework) is well-established (Cohen & Manion 1994) and now carries a good deal of intellectual clout. The problem in strategy research lies not in the fact of different methodological approaches, but rather in the lack of precision in the terminological apparatus it has employed. This may not altogether surprising, for (as I think Peter Medawar pointed out) there appears to be a pattern according to which concepts which migrate 'downhill' from more rigorous fields like psychology or cybernetics to somewhat less rigorous fields like education may retain little more than their labels.

Some definitions of strategy

Two of the more detailed attempts at clarifying the concept of strategy are discussed below. These are taken from Kirby 1988 and Faerch & Kasper 1983, as amplified by Bialystok 1984. Kirby (op.cit.:230) draws a distinction (here in relation to reading) between *strategies*, *styles* and *skills*. Skills are defined as "existing cognitive routines for performing specified tasks", while strategies are "the means of selecting, combining or redesigning those cognitive routines". Where a learner habitually makes use of a set or sets of similar or related strategies, we can talk of his or her *style.*

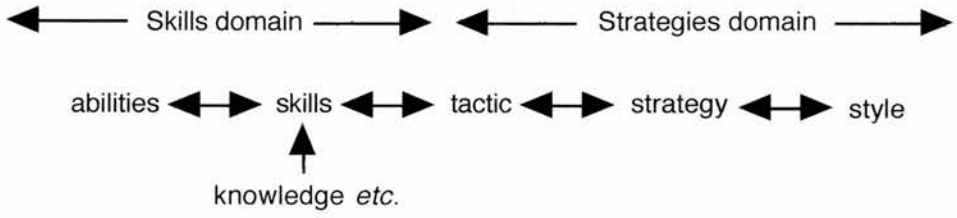

Figure 3.1: The relationship of skills and strategies (after Kirby 1988:230)

The skill and strategy domains, Kirby tells us, cannot exist without one another. There is a "constant interaction" between them, in that strategies condition the skills which will be brought to a given task, and skills influence which strategies are likely to be employed in task-completion. The strategies selected are a function of what the task-taker "expects" will help complete the task, and this is in turn a function of what the task-taker knows about the task and about her own skills, i.e. her 'metacognitive knowledge'.

Kirby offers a few concrete examples of skills vs. strategies. Reading *skills* include recognising words from features like their shape or structure, and constructing inferences, while reading *strategies* include deciding [sic] to use given features to recognise words, deciding to draw inferences etc. (One may decide to infer from this that for Kirby strategies are not necessarily conscious.) Again, however, (op.cit.:264) Kirby makes clear that:

> "...there is no firm dividing line between skills and strategies as categories of cognitive activity."

and notes that what may begin in a learner reader, whether in the L1 or L2, as strategy is likely to 'automatize' into a skill. For Kirby, skills are essentially "automated processes" carried out at a subconscious level, while strategies are more or less conscious activities. Given Kirby's caveat, above, it seems quite reasonable for Alderson 2000 (:310) to question the re-labelling as strategies of what have traditionally been regarded as skills (such as inferring word-meaning) but as with his critique of the looseness of the term 'cloze', the odds may be against him.

Ostensive definition

Kirby's attempt to define strategies in contrast to skills can be contrasted with attempts at *ostensive* definition (Bialystok 1984) exemplified by the many lists and taxonomies of strategies generated by researchers in the field (cf Block 1986; Cohen 1998) In this procedure strategies are first classified according to their putative function (learning, communication etc.) and then exemplars of each

functional class are documented. It is at this stage that the "conflicting results" noted by Skehan are most apparent. As Bialystok points out, however, even these functional classes are far from watertight, with "numerous overlaps" (op.cit.:38) as, for example, when communication strategies function as learning strategies and vice versa.

The difficulty of defining strategy, as it were, contrastively–theoretically may have encouraged the many attempts at ostensive definition. But there is another route by which a definition of strategy may be attempted: that of *defining features*. In her discussion of Faerch & Kasper 1983, Bialystok 1984 adds the feature of *intentionality* to the former writers' necessary or defining features of *problematicity* and *consciousness*.

Defining features

Bialystok's 'defining feature' of *problematicity* relates to the idea that strategies are 'triggered' by the occurrence of some problem in learning, comprehension or production. A second defining feature, *consciousness*, refers to the language-user's probable awareness (though on this point Faerch & Kasper are cautious, noting that the degree of consciousness of strategy use may vary among users and from situation to situation) of how a given strategy matches the problem in hand. Bialystok's notion of *intentionality* has to do with the user's control over his or her choice of strategies–i.e. some sort of deliberate selection among strategies is involved.

Why Bialystok sets up this third dimension of intentionality is not altogether clear, but it may reflect the idea that a subject may be conscious of having used a strategy without having chosen to do so; intentionality would then be a separate dimension. For Faerch & Kasper the boundary between unconscious and/or uncontrolled *processes*, on the one hand, and potentially conscious/potentially controlled *strategies* is blurred. They are rather more circumspect about the role of consciousness than is Bialystok herself, for whom the trait appears to be criterial. Two quotations will suffice to illustrate this point. (Communicative) strategies are seen by Faerch & Kasper as:

> "*potentially* conscious plans for solving what, to an individual, presents itself as a problem in reaching a particular communicative goal."

(Faerch & Kasper 1983 cited in Poulisse et al 1987; italics mine) while (learning) strategies are defined in Bialystock 1978 as

> "optional means for exploiting available information to improve competence in a second language"

Bialystok's specification of 'optionality' in this context appears to imply conscious control. The clear overlap between communication and learning strategies is reflected in these and other attempts at definition. Indeed, Faerch & Kasper's definition of communication strategies has been cited in studies of strategies of ostensibly different kinds. In short, there appears to be no real consensus as to how strategies should be categorised or even what feature(s) identify them as such. In the absence of any overall framework, individual

researchers have tended to set up their own identification criteria and lists of strategies; the negative implications of this for cross-study comparability are plain.

3.11 Influences on strategy use

Oxford 1990 lists a number of factors affecting strategy use. *Motivation* appears to be an important influence, in that less motivated learners appear to use fewer strategies, and these less often. *Cultural background* may also play a role in that certain strategies may be alien to particular cultures, e.g. (my own example) the reluctance of Japanese learners to put questions to teachers. The *task-type* is clearly likely to influence which strategies are employed, and *learning style* may play a part too. Oxford cites *age* as a potential influence on strategy choice, but does not go into detail. One probably crucial factor, however, is the *learner's proficiency level in the target language*. A learner-reader still struggling to recognise basic sentence forms would, for instance, be less likely to make use of skimming as a reading strategy. In connection to Oxford's cultural background factor I would add that of *the nature of the L1*. Japanese, for example, possesses a great number of homophones, some of which may be distinguished in the absence of sufficient context (e.g. within a name) only with reference by the *kanji* characters used to write them. In learning their L1 students develop a reliance on their dictionaries to guide their choice of characters in writing, and this 'reluctance to trust the ear' appears to spill over into their L2 lacquisition.

3.12 Categorisations of strategies

However a definition of strategy is reached, classification and exemplification will at some stage be necessary. Oxford notes that, to date, there are around twenty-five rival systems of classifying learner strategies, and that these tend to be based on one of five paradigms. Naiman *et al* 1978 base their systems on an analysis of the behaviours of successful language learners. O'Malley & Chamot 1990 base theirs on a list of psychological functions common to language users. Bialystok 1978 makes use of linguistic strategies, while Cohen 1998 bases his on language skills. Oxford herself (1990) uses the criterion of learner-type. As she points out, there exists a pressing need for some established and accepted system of classification.

Another feature of much current research inimical to cross-comparison is the variety of stimulus tasks set, and the strategies identified seem to be strongly affected by the nature of the elicitation task and on the theoretical and *pre*theoretical decisions researchers have made. Oxford's figure of twenty-five taxonomies notwithstanding, there can appear to be almost as many categorisations of strategies as there have been studies of them. Classification schemes for strategies range from plausible but uninformative binary systems to the thirty-four or so 'processing activities' identified by Simons & Lodewijks 1988 (cited in Pressley & Afflerbach 1995.) Halliday remarks somewhere–in reference to schemes of language function–that in the absence of objective classification criteria there is little to choose among rival taxonomies. Much the same must

apply to classifications of strategies, so that researchers are to some degree (but cf. Faerch 1984:69) encouraged to construct *ad hoc* category schemes for their data. Some classifications relevant to my own focus on cloze processing are discussed in a later section.

Some dichotomous schemes

Johnson 1983 identifies two main types of comprehension strategy, those which aid the reader in creating a framework of meaning in order to make sense of the text, and those which monitor the comprehension *process*–triggering compensatory actions when understanding breaks down. Hosenfeld 1977 sees the key distinction as between "main meaning line" and "word-solving" strategies. The first of these appears to be essentially the same as the first of Johnson's strategies above, while the second of Hosenfeld's strategies seems to be one type of processing 'repair'–commonly known as lexical inferencing. Feldmann & Stemmer 1987 distinguish 'recall' and 'evaluation' strategies . The former aid recovery of (in their case study) C-test deletions, while the latter are used in checking the recovered items. How clearly this distinction between recovery, or generation on the one hand, and evaluation on the other can be maintained in practice is moot, but I accept that the latter category is a psychologically real one identifiable in informant data.

More elaborate categorizations

Ballstaedt & Mandel 1984 employ the label 'elaborations' to refer to the class of (inferencing) comprehension strategies in which they are interested. Either unable

or unwilling to attempt a scheme of strategies, they look instead at whether a subject's "learning success can be predicted [from] the number of [elaborations]" his or her protocol contains. Ballstaedt & Mandel analysed protocols according to three criteria:

1. the number of words
2. the number of main ideas ("new, self contained ideas" seen as equivalent to "meaning units")
3. the number of "foci of attention", which are demarcated according to syntactic and/or phonological criteria

Ballstaedt & Mandel found a high correlation between these measures, including the rather crude measure of word-count, and learning/comprehension success. For some limited purposes, then, sophisticated analysis of protocols may be superfluous.

Block 1986

Block's 1986 survey of research into reading strategies, while a valuable guide to work-to-date, underlines the diversity among the studies surveyed in terms of age and grade-level of subjects, in reading materials and tasks employed, and in the instruments used to measure comprehension and isolate strategy–use. These diversities combine to make cross-study comparison highly problematic.

In addition to surveying work-to-date, Block reports on her own study of native and non-native speaker reading strategies. The resulting catalogue of strategies identified sets up a binary division of these into five 'local linguistic' and ten

'general comprehension' strategies, with the latter further divided into strategies of 'comprehension gathering' and those of 'comprehension monitoring' (cf. Johnson 1983). Given its clear influence on subsequent studies, Block's scheme is worth looking at it detail. Using this scheme of strategies to analyse a set of (immediate retrospection) reading-task verbal report protocols produced by native speaker and non-native speaker subjects, Block found an inter-rater reliability of over 80%. This suggests that the above scheme is relatively straightforward in application.

Block's categorisation of strategies appears to follow those of Johnson, Hosenfeld and others in distinguishing what we might characterise as strategies more narrowly focusing on the text itself, and those which integrate the reader's background knowledge with the text's information content. It is difficult not to see a parallel between this distinction and that between 'bottom-up' and 'top down' reading strategies (Carrell 1987). Block's scheme has been the basis of several subsequent taxonomies (cf. Nevo 1989; Trollope 1995) and this may be due to the fact that her categorisation is a synthesis of–and is hence seen as an improvement on–a number of previous schemes. Block's taxonomy also appears to have what used to be called 'high face validity' in that a typical verbal report protocol will contain a number of items readily categorisable under her scheme. In retrospective interviews which I conducted with two native-speaker respondents without any previous experience of think-aloud, both were able to map stretches of their audio-taped data onto Block's 1986 categories. While other schemes of

categorisation (cf. O'Malley & Chamot 1980) are arguably more comprehensive or more detailed, Block's appears to be fairly practical. If cloze-type tests are indeed able to tap much the same processes as are activated during reading (Hinofotis & Snow 1980) then categorisations of reading strategies like this may be relevant for the analysis of verbal reports of cloze-processing.

3.13 What do we know about test-taking strategies?

> "What a test of reading tests is not simply what its constructors say it tests, nor what a set of judges considers it to test. It must surely and crucially relate to what happens inside a test-taker's head when he or she responds to an item." (Alderson, 1990:478)

Alderson's implicit defence of introspective data, above, links the notion of strategy with the area of test-taking. In recent years more studies have begun to look at the processes of test-taking, although the number of studies published remains small in comparison to the quantitative/statistical data on scores. In Israel, the USA and in Germany research into test-taking strategies via verbal report data has been the subject of a number of theses and dissertations. Studies discussed in Cohen 1998 found that weaker test-takers used "low-level" word-focused strategies–referring to dictionaries and translated extensively–while contextual guessing was largely the preserve of stronger test-takers, and readers' background knowledge about a topic influenced the extent to which they were able or willing to make inferences.

Studies of cue-uptake in text processing

A first step in establishing a taxonomy of cloze cue-uptake behaviours must be to look at what has already been attempted in this field. Few, in fact, of the literally thousands of studies of cloze procedure have looked at subjects' test-taking processes. Even fewer of these have been published, so that much of the work that has been done is to be found (eventually) in dissertations or internal research reports.

I can look here only at the most relevant studies encountered so far, focusing–if not on cloze-processing itself–then on tasks whose cognitive demands appear to overlap those set by cloze. i.e. lexical inferencing tasks and C-Test. The studies touched on, or looked at in more depth, here are Markham 1987, Kesar 1990, Manghubai 1990, Allan 1992 and Trollope 1995–all studies of cloze or cloze-like tasks. Block's 1986 and 1992 papers on reading behaviours–assumed to be of relevance for cloze and lexical inferencing–are discussed next, followed by Stemmer & Feldmann's 1987 study of C-Test strategies. Ames' interesting and ambitious 1966 study of lexical inferencing is accompanied by a look at Haastrup's 1991 study of the same task.

'Strategies' versus 'cues'

I would stress here the functional identity of many categories of *cues-to-recovery-of-meaning* and of *strategies-of recovering-meaning*. For example, we find that many taxonomies of cues-to-meaning have a category along the lines of *cues-within-local-context (same clause/sentence)*. This seems to me to

be just another way of expressing the strategy *search local context (same clause/sentence).* Meaning-cues, after all, are only realised through an informant's exploitation of them. (It would not be unreasonable to talk of a *skill* of searching local content, but terminology seems to have a dynamic of its own.)

Two main modes of classifying context cues

There are two main criteria by which contextual cues are classified in the studies discussed here. These are, as briefly discussed in this chapter, firstly by their formal or semantic relationship to the target item, and secondly by their physical distance within in the text from the target item. While in the former we might find cues classified into synonyms of the target, co-elements in a fixed expression, superordinate terms, etc., in the latter we might expect to see cues classified according to whether they are located within the same clause as the cue, within the same paragraph and so on. Schemes for categorising cues may utilise either or both of these criteria according to their purposes and, given the ongoing debate over how well cloze taps constraint at the level of local or global passage context, it is not surprising that cue–target distance has been a major concern in studies of cloze processing.

Here I will look critically at the various schemes of cue 'type' and cue 'distance' in the literature, and attempt to see what they might contribute to a workable and plausible classification scheme for my own purposes. I look first at a number of small-scale studies which have attempted to uncover the strategies used by

subjects in rational-deletion cloze tasks.

Markham 1987 and rational cloze processing

Markham 1987 used retrospective interviewing to access the (rational deletion) cloze processing strategies of ESL and NS subjects. His apparent focus, however, appears to have been on subjects' use of intrasentential and inter-sentential cues in recovering deletions. Markham's study, then, is essentially to do with target-cue distance, and the question of to what degree cloze taps global comprehension.

Introspection may offer an alternative or supplement to the procedures such as comparisons of scrambled vs. sequential text (cf. Oller & Jonz 1994) conventionally used to examine this question, but Markham's paper exhibits–for our purposes here–a number of shortcomings. For one thing, we are not told whether ESL informants reported in their L1 or L2 (My own experience of data-gathering suggests this is of primary importance.) nor are we told what 'memory supports', if any, were made available during the retrospective interviews. These took place as long as six days after administration of the cloze task, and such a delay between task and retrospection raises the question of how well Markham's informants can have recalled their own processing behaviours. This is especially problematic if they were allowed no memory supports such as the opportunity to audit their recorded task protocols. Again, experience suggests that provision of memory support can be decisive in (a) making recall possible at all and (b) preventing or identifying informants' reconstructions.

That said, Markham's 1987 study produced the interesting finding that there were "no significant differences" between the cue-uptake strategies of ESL and NS subjects, except that ESL subjects "actually made slightly greater use of intersentential cues than natives" (:306) in all categories of target item except 'substantives' (i.e. lexical-item deletions.) Markham goes on to suggest that "the complexity of trying to delete only those [...] words that truly cause students to focus on intersentential constraints" makes rational cloze an unlikely means of testing global comprehension. Cloze test-takers' introspections about their processing may, however, offer a more efficient way of uncovering just when, and under what circumstances, intersentential cues are in fact activated.

Three types of cue

Markham 1987 separates cues aiding rational deletion (content word) cloze into three categories: 'intrasentential' (i.e. within the same sentence as the target), intersentential (i.e. in another sentence), and 'pragmatic', i.e. depending on extratextual knowledge. The intra-/intersentential distinction relies on the widely-accepted finding that the sentence or clause functions as an information unit, in that integration of new information into the existing representation of text meaning tends to take place at the ends of sentences (Taft 1991.) This intra-/inter-sentential distinction need have little to do with physical text distance between target and cue(s), however. As the lexical inference example (my own) below demonstrates (target in italics; cue underlined), target item and meaning cue may be adjacent in the text, yet still be divided by a sentence boundary:

> One of the most persistent beliefs, in the countryside at least is in the influence of one's *maalim*. This first adult (other than a female blood relative) whom the mother meets on leaving the hut in which she gave birth is believed to have a great influence on the course of her child's early life. It is common in rural areas to see a new mother lurking , with eyes downcast, just inside the doorway of her family's hut, while her own mother or sister goes to fetch a suitable *maalim*.

Markham's sole concern with different processing behaviours lies in the ratio of intra- to intersentential cues utilised by cloze test-takers. His findings suggest that targeting content words does not in fact require test-takers to process the text at a global level. "Generally speaking", he tells us, "it does not appear necessary to pay attention to the global cues in order to complete the deletion." As always in this aspect of the cloze debate, much depends on what one means by 'generally speaking.'

Kesar 1990: reading processes in rational deletion cloze

Kesar 1990 investigated the reading processes used by subjects completing rational-deletion cloze tasks. Keisar claims to have been able to recruit subjects with no previous exposure to cloze-tasks, and in her data she identifies between twenty-six and thirty "cognitive and meta-cognitive" reading strategies. I am unable to describe the individual strategies in any detail due to the fairly skeletal nature of the abstract (and my non-existent command of Hebrew, the language of publication) but one finding Kesar notes is that (sub)sentence-level ('micro' level) strategies were the most widely used, by weaker subjects as well as stronger, while use of 'macro' level strategies tended to be the preserve of the latter group.

Stronger subjects, then, are able to "[see] the text...as a developing unit of meaning." This appears to sit well with Markham's findings, above.

Although I discuss schemes of target-cue distance later in this chapter, it may be as well here to outline my limited and second-hand picture of how Kesar sees this aspect of cloze processing. The basic division appears to be between 'micro-level' and 'macro-level'. The former embraces word-level/part-of-sentence level, and the latter five so-called discourse levels: intersentential; whole-text; extratextual; "metacognitive"; and "other." My interpretation is that the 'intersentential' category is roughly equivalent to Sasaki 1993's "across sentence, within paragraph", in other words designed to allow for cue-seeking and uptake beyond local context, but not truly at global passage level. The "metacognitive" level appears to be a behavioural category rather than a measure of target-cue distance, and (if I have understood it correctly) indicates that an informant made some comment about some aspect of her processing of the cloze passage or the verbal-report task. "Other" appears to be a catch-all for uncategorizable behaviours or events.

Accessing meaning in cloze: Manghubai 1990

Mangubhai 1990 is more directly relevant to my own work insofar as he used think-aloud procedures to access the cloze test-taking behaviour of six Fijian secondary-school-age subjects in a purely concurrent think-aloud procedure. From the verbal report data these subjects produced, Mangubhai assembled an *a*

posteriori list of strategies. Although only a few pages in length, Mangubhai's study is one of the very few to apply a concurrent instrument verbal report to the study of cloze processing. Mangubhai's taxonomy of strategies is discussed below, but–if only to emphasise that much research remains to be done in this field–certain limitations inherent in his study may be noted in passing.

Firstly, no retrospective comments appear to have been elicited from subjects. Subject-retrospection about previous think-aloud tasks has been criticised (Ericsson & Simon 1984, 1993) but it has also been argued (ibid.; Haastrup 1991) that retrospective interviewing can provide a degree of 'control' on the researcher's (and even the informant's) conclusions about processing behaviour during the stimulus task.

Moreover, Mangubhai makes no attempt to correlate his own taxonomy of subject behaviours with those of other researchers. It is thought desirable (Faerch & Kasper 1987) to at least attempt to correlate a new study's behavioural categories with those utilised in previous research so that cross-study conclusions may be drawn. It can, however, equally be argued that in order to minimise 'contamination' from previous studies–whose aims, conditions and procedures may well be quite different–one should at least initially base any analysis purely on one's own data. What seems clear is that at some stage cross-study comparison of strategies should take place, if only to ascertain the degree to which typologies of strategy can or cannot be made to coincide. In fairness, however, Mangubhai's

paper is only a few pages in length, and there is no indication from the bibliography that the author was aware of what little other research has been done in this area. Moreover, as Pressley & Afflerbach 1995 point out, by no means all researchers in the areas of processing strategy and introspective data describe their categories in a way that makes cross-study comparison viable. Mangubhai's "category coding scheme" is set out below.

Task Strategies

A Looks at immediate context before generating a word
B Looks at the larger context before generating a word
C Looks at immediate context after generating a word
D Looks at the larger context after generating a word
E Generates a word but does not confirm its suitability for the gap
F Generates a word and evaluates its correctness in the gap
G Gives up after one or more attempts to fill the gap to come to it later
H Gives up after attempting to evaluate the correctness of a word and accepts it

Cognitive Strategies

A Refers to prior knowledge
B Rephrases the sentence in order to generate a word
C Repeats a word/ words so as to retrieve an associated word from memory
D Seeks collocation or gets the word through collocation
E Serially generates words of one class
F Generates randomly and rejects word on syntactic or semantic grounds
G Changes consciously or unconsciously word(s) of the passage in order to allow the generated word to fit
H Analyses the passage, using prior and contextual knowledge in order to generate the word

Judgement of Correctness (i.e. evaluation strategies)

A Judges by 'feel'–sounds correct

B Uses rule(s)

C Uses knowledge of the contents of the passage

D Cannot tell

Automatic Processing

A Automatically generates word that is correct

B Automatically generates word that is incorrect

Comments on Manghubai's scheme

Given the lack of any transcribed data in Manghubai's paper, it is difficult to assess how accurately his categories match the subjects' behaviours. We may, however, comment on how plausible and coherent his scheme appears to be, and how useful it might be for future research. Looking first at Manghubai's 'task strategies', my intuition is that A, B, C and D correspond to behaviours I have observed in my own trials. A and B may be seen as 'generating' behaviours, while C and D represent 'checking' stages. Mangubhai's strategies E and F are more problematic, as both

E Generates a word but does not confirm its suitability for the gap

F Generates a word and evaluates its correctness in the gap

appear to me to conflate two operations into one.

'Task strategies' G and H seem more plausible, as do 'cognitive strategies' A through C. D may be more plausible if we drop the "seeks collocation" element, and E through H are again intuitively acceptable. Similarly, all four of Mangubhai's 'evaluation' strategies, and both of his 'automatic' strategies seem

to correspond to observed behaviours. My only caveat here is that the strategy of [getting a] word through collocation may largely be an automatic process (Taft 1991) although it may reach the level of attention and thus be reportable.

The 'before' and 'after' problem

In the light of close examination of verbal report protocols, I would question Mangubhai's neat 'before' and 'after' distinctions between strategies A and C, and B and D. As I have noted elsewhere, in the absence of REM monitoring equipment, it is hard to tell with any precision just where in a text a subject's eye is falling at any given time. We might be on safer ground if we substitute the word 'verbalise' for 'generate', as the only thing which can be located precisely in time in studies such as this is the point at which cognitions are externalised.

It is not clear to me what Manghubai means by his strategy

E Serially generates words of one class

It is common for subjects to generate a series of, say, verbs or nouns in an apparent attempt to arrive by 'feel' or by 'sound' at the optimal choice of filler. As important as common class membership here may be common membership of a semantic field or set, as in the example below from my own data:

"some official reports.. books.. informations [sic].. lists.. ranking... hmm.. hmm uh.. mm.. take report.. reports"
(GL1 think aloud Helga)

As this informant later made clear–an indication of the value of post-task interviewing–she was seeking the best choice of word with the German meaning of '*Unterlagen*' (official documents), and her criterion for selection seems to be something along the lines of 'belonging to the lexical area of bureaucracy'. Whatever limitations it may have, Mangubhai's paper at least moves beyond a narrow concern with target-cue distance relationships to look at some of the wider behaviours and processes cloze test-takers may engage in.

Allan 1992: Strategies in solving rational deletion cloze

While Kesar appears to be neutral on the question on whether cloze represents a valid method of testing reading comprehension, Allan 1992 suggests that his findings put a question mark over cloze as a reading comprehension measure. Allan identifies three classes of strategy: (1) those used in "approaching the task for the first time", (2) those used in recovering individual items, and (3) those involved in evaluating individual recoveries and success on the test as a whole. Allan claims that the most common "approach to the exercise as a whole" (by which I take him to mean his first class of strategy, i.e. approaching the task for the first time) was to read the complete text. This behaviour was often applied when "major difficulties were encountered" in recovering particular items, although at times only the paragraph containing the item was read. It appears, then, that an 'approach' strategy can also be applied not just to the task passage as a whole, but also to individual items as and when necessary. This seems to some extent to conflate, if you like, action and reaction.

Moreover, although initial reading of the full text is commonly recommended or specified in cloze and cloze-like task rubrics, Allan's finding that pre-reading of the entire passage was his informants' most common approach to the task. It is not clear how much freedom Allan's informants had to adopt another approach, but Cohen 1998 notes that studies suggest that typically only around one quarter of cloze test-takers actually read the full passage before beginning to recover deletions. In my own data collection to date few informants have not read at least a sentences or two of the passage before beginning recovery, while of those who read more extensively (again, only a minority appears to have read the passage in full it should be noted that test-takers may be recovering 'easy' deletions while ostensibly 'just reading', and 'reading' perhaps at quite high speed. This is evidenced in cases when a test-taker inserts filler-words at speed more or less immediately after she has ceased to 'read aloud' the passage span in which these were located. It is difficult to believe that these 'reading aloud' stages contained nothing but reading nearby extant text, so that pre-reading in cloze processing may be a less discrete or separable behaviour than some writers assume.

One suggestion may be introduced here: an alternative to postulating separate 'classes' of strategy is to identify strategies which are statistically more or less likely to be applied in particular contexts, such as when 'approaching' a text, recovering individual items etc. Thus, while an initial reading of the full passage is undoubtedly a common 'encounter' behaviour, it need not be confined to a first approach to the task. 'Recovery' strategies identified by Allan (listed

verbatim here) were:

(1) reading the context surrounding a gapped item
(2) reading the context up to a gap substituting it with the word 'what'
(3) skim reading the whole passage
(4) repeating the word preceding the blank
(5) skipping items
(6) searching for clues later in the passage, a "test-wise" strategy

The first of these strategies appears to equate to reliance on the 'immediate context' of the item, while strategy (2) in Allan's scheme appears to me to contrast with (6), i.e. it reflects the operation of seeking clues in the text up to and including the deletion, but not beyond it. Allan himself sees the 'what'–insertion of strategy (2) as an attempt to 'repair' the text temporarily, and notes that it may applied to several gaps in succession. He also notes (as does Bachman 1990) that subjects may fill a gap 'provisionally', often making explicit in their protocols that they have limited confidence in the word used.

The word substituted for the deleted item in a strategy like (2) is, I would suggest, only important insofar as L1 equivalents (such as German "*irgendwas*", Japanese "*nantoka*") or even sounds ("*dududu*", "*lalala*") can be distinguished in terms of the extent to which they represent what Allan 1992 calls a 'repair' (and what I term a 'running' or provisional filler') as opposed to an index of perceived difficulty, uncertainty, etc. This is marginally possible, but as it may involve extremely close attention to the informant's intonation as she utters the filler word or sound (thus calling for a very clear audio-recording) and interruptive questions

to obtain her immediate retrospection, it is hardly practical on any scale. It is less plausible, however, that any 'real' words–i.e. other than interrogatives or nonsense items–employed as provisional fillers are selected arbitrarily. The relationship of provisional to final recoveries may well merit examination, and the former are very often quite clearly marked by a rising intonation pattern that strongly implies the items tentative nature in the informant's mind.

It is tempting to see Allan's third strategy, "skim reading the whole passage", as a kind of fallback behaviour indicating marked difficulty in recovering the item, or at least an awareness that cues are not to be found in the local context. If this hypothesis is valid, we might expect full passage skimming to follow attempts at recovery through more local text searches. Similarly, "skipping" an item (5) typically represents (according to my informants) some considerable difficulty in recovering it. This may indicate that the informant cannot think of a suitable filler, or is unable to choose among candidate fillers. It can, however, also mean that she wishes to move on to another–typically a subsequent–item in the hope of finding there some clue which will help her fill the skipped gap of choose among her candidate recoveries.

Allan's fourth strategy, that of "repeating the word preceding the blank", might usefully be extended to cover the repetition of a text string preceding the deletion. Behaviour like this occurs frequently in verbal report protocols, and Allan makes the intuitively plausible suggestion that this represents a "process of association"

in which subjects attempt to generate words which 'sound' like an appropriate continuation of the word sequence. Given the heavy emphasis on rote-learning in school English courses in many Asian countries, it is quite possible that an observation by one of my informants applied not only within the Japanese context. This informant claimed that she was not so much using sound alone in the following protocol extract as also trying to visualize textbook or classroom-printout contexts in which she might have seen the appropriate "phrase":

"...of the full moon was set set set.. was set SOMETHING as a holiday.. was set set set set.. as set as.. as SOMETHING as was set mmm? was [...] it was..SOMETHING.. set ah set ummm set set set full moon was set off? set on as a holiday?.."
(JL1 think aloud Kazue)

The final strategy in Allan's scheme involves (if I understand Allan correctly) searching the passage following the deletion as well as before it. Quite why he labels this a 'test-wise' strategy is unclear, for searching the cotext for clues to meaning is a behaviour observed in lexical inferencing (Aizawa 1998), a task which quite may quite tightly overlap authentic reading. I have also found it to occur widely in a classroom reading activity I developed some years ago in which readers who notice (not all do) an unexplained lexical item within the passage search (often extensively) for a paraphrase or explanation of its meaning. To label strategy (6), as Allan does, seems to imply that it is a peculiarly test-oriented behaviour, which I do not think it can be unless 'test-wiseness' is reduced to something like 'if *x* doesn't work, try *y*.'

Cloze-processing strategies in Trollope 1995

Trollope's study is another of the few focusing on cloze processing *per se*, albeit partly in the context of on-screen presentation, and it is worth looking at his taxonomy of "on- and off-line reading strategies". Trollope claims to have used a modified rational-deletion technique, in which he deleted every tenth passage word where this was a content word. Where this was not the case, he deleted the closest content word before continuing, and in this way he hoped to achieve an average deletion ratio of 1:10. This blend of fixed-ratio and rational deletion may reflect the lack of a criterion for choosing among content words to delete. The examples given here are taken from Trollope, except where my own data appeared to offer a clearer illustration.

Trollope's categories are briefly described below, and it should be noted that these include a number of strategies which pertain only to on-screen cloze processing, as well as some relevant only to test contexts in which dictionary reference is allowed. For reasons of length I have omitted the above strategies, which are not relevant to my own data-gathering situation and are, I think, not the normal contexts in which second-language learners or testees encounter cloze or cloze-like tasks. I discuss here, then only those pertinent to on-paper presentation of cloze, *without* dictionary use.

[1] ANTICIPATE CONTENT: Here the reader explicitly anticipates what will appear in upcoming passage text

[2] APPEAL TO RESEARCHER FOR ASSISTANCE: Self explanatory,

although Trollope is not explicit about whether this behaviour was explicitly sanctioned, or whether and how much assistance was given to informants who requested it.

[3] APPEAL FOR ASSISTANCE/ASK OPINION OF PARTNER: One reader asks her partner for assistance, or for her interpretation. According to Trollope, this is usually triggered by comprehension difficulty.

[4] CORRECT BEHAVIOUR: The reader notices that an assumption, interpretation, or paraphrase is incorrect and makes the necessary revisions. Says Trollope "This is a combination of strategies of integration and monitoring, since the reader must both connect new information with old and evaluate understanding."

[5] COMMENT ON BEHAVIOUR OR PROCESS: In such a 'metacomment' the reader makes an explicit reference to her cognitive processing, expresses some awareness of comprehension success or failure etc..

[6] TRY TO UNDERSTAND THE TEXT AS A WHOLE (GLOBAL UNDERSTANDING): The reader tries to construct or (re)state a global meaning for the text in response to a local level comprehension problem.

[7] IGNORE WORD/SENTENCE/CLAUSE: The processor explicitly states that a given string is not necessary to the task, or that it will be returned to at a later stage.

[8] INTEGRATE INFORMATION: New information is related to information acquired from prior text. Trollope does not make clear whether atextual prior knowledge is also a candidate for integration.

[8] INTERPRET THE TEXT: The reader makes an inference, draws a conclusion, or forms a hypothesis about the content–or attempts to do so.

[10] MONITOR COMPREHENSION: The reader explicitly comments on her understanding of (some part of) the passage.

[11] PROCEDURAL MOVE: One or both readers in a DC task comment on their task procedure.

"While you're looking up that I'll go on to the next one."

[12] QUESTION INFORMATION IN THE TEXT: The reader questions the importance or veracity of some aspect of the text.

"Why is [baby talk among adults] usually limited to lovers?"

[13] RECOGNISE TEXT STRUCTURE: (Trollope) "The reader distinguishes between main points and supporting details or discusses the purpose of information. Responses occurred in the extensive mode.

"Because sometimes they have a string of umm nouns following the first head noun." "

[14] REACT TO THE TEXT: The reader makes an explicit or implicit affective response to text content.

"So they...the whole year has to be named after one person? Hmph."

[15] REFER TO TITLE: The reader tries to gain some insight into text meaning by reference to the title.

"Well the title says the voice of umm.. what's maître?"

[16] SKIM FOR GIST: The reader skim reads the text for its overall sense.

[17] REFER TO THE SOURCE OF THE ARTICLE: The reader asks the researcher where the article came from (i.e. newspaper, magazine, etc.)

[18] USE GENERAL KNOWLEDGE AND ASSOCIATIONS: The reader uses extratextual knowledge to perform any of these functions: to clarify or expand on text content; to evaluate the veracity of content; to react to content.

Trollope goes on to list a set of "local strategies" (19–30); the inference from this must be that the strategies above are somehow more 'global'.

[19] ABANDONMENT: The reader "gives up" implying or stating that she will make no further attempt to recover the item in question.

[20] GUESS THE WORD NEEDED OR MEANING OF WORD: Trollope's gloss of this strategy may be open to question: "The reader proposes a word or interpretation without taking into account the context." Of Trollope's two criteria for 'guessing', the informant's explicit claim that *x* is a guess is perhaps more satisfactory than that "the proposed word or interpretation was incoherent."

[21] PARAPHRASE IN L1: "The reader rephrases content using different words, but with the same sense." According to Trollope, the strategy is used "to aid understanding, to consolidate ideas, or to introduce a reaction."

[22] PARAPHRASE IN L2: As above, but in the second language.

[23] QUESTION MEANING OF CLAUSE OR SENTENCE: The reader states that she does not understand the meaning of a text string.

[24] QUESTION MEANING OF A WORD: As above, but at word-level

[25] REFER TO THE GRAMMATICAL CONSTITUENTS OF A WORD: Says Trollope, "[t]he reader analyses the prefixes/suffixes or inflexions, etc. of a word to aid comprehension or test hypotheses. This includes analysing punctuation to aid comprehension or test hypotheses." Obviously, in a cloze task, the word analysed must be a cotextual item.

[26] REREAD: "The reader rereads a portion of the text either aloud or silently." Trollope sees this behaviour as indicating a lack of understanding, but acknowledges that it may also be used as a

means of "[giving] the reader time to reflect on the content."

[27] SOLVE VOCABULARY PROBLEM: "The reader uses context, a synonym, or some other word-solving behaviour to understand a particular word."

[28] TEST HYPOTHESIS: The reader puts in a filler which she explicitly states she is not sure about.

[29] TRANSLATE SENTENCE OR CLAUSE: The reader translates the sentence/clause to gain understanding of the text.

[30] TRANSLATE WORD WITHOUT DICTIONARY: The reader translates a text word via word attack strategies such as recognising cognates, guessing from the context, etc.

Block 1986

Block 1986–among the most widely cited studies of learners' strategic behaviour–used a think-aloud procedure to examine the reading strategies of L2 readers' processing a freshman-level reading passage. Block proposed a set of descriptors of "local strategies" geared towards the understanding of "specific linguistic units" (see also section 4.4.3.) Although only one of these–Block's strategy number 15 ("Solve vocabulary problem: The reader uses context, a synonym or some other word-solving behaviour to understand a particular word")–appears directly relevant to the cloze task, many of the strategies listed on page 492 of her paper seem to correspond loosely to behaviours observed in my own data from cloze test-takers These include:

Integrate information (Strategy 3, below)
Use general knowledge and associations (Strategy 6, below)
Comment on behaviour (Strategy 7)
Monitor comprehension (Strategy 8)
React to the text (Strategy 10)

Paraphrase text (Strategy 11)

Reread span of text (Strategy 12)

Question meaning of text segment (Strategies 12, 13)

Block's 1986 scheme of strategies is set out in full here. Her 'general comprehension' strategies consist of:

1. Anticipate content (predict upcoming information)
2. Recognise text structure (distinguish main from subordinate points)
3. Integrate information (connect new information with old)
4. Question information (question veracity or significance of information in text)
5. Interpret text (draw inferences, conclusions, hypotheses)
6. Use general knowledge & associations (to evaluate, extend, explain or react to information in text)
7. Comment on behaviour or process (describe affective responses, strategies, problems etc. In comprehension process)
8. Monitor comprehension (assess depth of understanding)
9. Correct behaviour (correct assumptions, interpretations in the light of textual information)
10. React to the text (react emotionally to information content)

while her 'local linguistic' strategies are:

11. Paraphrase (express the same meaning in different words to aid understanding, consolidate ideas etc.)
12. Reread (aloud or silently; usually reflects a lack of understanding, but can signal reflection on information content)
13. Question meaning of clause or sentence (indicates lack of comprehension of part of text)
14. Question meaning of word (indicates lack of comprehension of word)
15. Solve vocabulary problem (use context or synonym to establish meaning of a particular word)

Block 1992: A more flexible model?

Block 1992 identifies the following categories of event in subjects' verbal reports of pronoun-referent identification and lexical inference during reading. These categories may be combined to describe data events:

(1) PLAN refers to a subject statement of how she is proceeding with the task.
(2) PROBLEM RECOGNITION is self-explanatory (and is less directly relevant to cloze processing, where the 'problem' is provided.)
(3) SOURCE IDENTIFICATION refers to subject's identification of relevant information.
(4) SOLVE refers to the subject's guess or conclusion about pronoun reference or word meaning.
(5) CHECK refers to the subject's actions in assessing the appropriateness of her solution.
(6) REVISE refers to the subject's actions in amending an existing solution.

Block's 1992 categories are clearly quite different in orientation from those of six years earlier, and appear to represent a move towards a more flexible list of (combinable) 'functional' operations. There may be, after all, limited benefit to be had from distinguishing among actions or behaviours which have much the same function in terms of textual processing. But this is not to neglect distinctions which are plausible and of value: SOURCE IDENTIFICATION, for example, might usefully be further broken down into uptake of intratextual and extratextual cues to meaning. Block's revision of her 1986 list of behaviours seems positive, however. On the basis of my own data, I would question how separable some of

Block's 1986 strategies really are: a 'comment on behaviour' such as

"Umm...I HAVE TO READ THIS PART AGAIN.. it's not clear what..."
(G11 think aloud Manfred)

might easily be taken as representing her 'monitoring of comprehension' function.

Strategies of C-Test-taking

Feldmann & Stemmer 1987 is one of a number of recent studies focusing on the C-Test. This offshoot of cloze procedure was designed (Klein-Braley 1981) to remedy some of the perceived shortcomings of 'standard' cloze.

As mentioned in chapter 2, C-Test and cloze are both said to measure a test-taker's ability to make use of reduced-redundancy in interpreting a passage. There are grounds, then, for suspecting that test-takers may employ some of the same strategies or behaviours in recovering deleted individual words in cloze, or in completing mutilated words in the C-Test. I reproduce here only those of Feldmann & Stemmer's C-Test-taking strategies which may be salient in the processing of cloze passages, that is, those cues which do not rely on the presence of the first half of the mutilated C-Test item, and the morphological or phonetic cues it may carry. Strategies are numbered as they appear in Feldmann & Stemmer 1987, and are divided into what the authors label 'recall' and 'evaluation' strategies. The authors make clear that this scheme is provisional and quite possibly incomplete. Little data appears to have has been collected on the

frequency with which individual strategies were employed.

'Recall' strategies

"Recall strategies", Feldmann & Stemmer tell us, "are employed ... where the retrieval of an item is non-automatic." v(op.cit.:259) These recall strategies comprise:

[1] Recall by structural analysis
  [a] Syntax analysis (explicit reference to pronouns etc.)
  [b] Formal indicators (punctuation markers)

[3] Recall by repetition
  [b] of following word(s), sentence(s)
  [c] of text (rereading of text)

[4] Recall of past situation (e.g. previous exposure to item via reading or teaching)

[5] Recall by search for meaning
  [a] translation to L1
  [a1] translation of directly following words
  [a2] translation of directly preceding words
  [a3] translation of further co-text
  [b] looking for L1 meaning equivalent or use of hypothesised L1 meaning equivalent
  [c] looking for general meaning of text

[6] Recall by looking for external help (other tests etc.)

[7] Recall by substitution (provisional filler inserted, e.g. 'something'; '*dadada*' etc.)

[8] No identifiable strategy

The heuristic value of [8], above, and [4] below may be questionable: presumably the authors apply these when they feel sure that *something* is going on, but have no real idea what it night be. Although it is perhaps inevitable that the authors will label it thus, their 'recall by substitution' strategy ([7], above) appears in my own data to have more to do with preventing the informant's *lack* of recall from interfering with the rest of the task. It may represent a kind of place-holding behaviour, rather than recall as such.

'Evaluation' strategies

In Feldmann & Stemmer's view, evaluation strategies "are used in order to check on the appropriateness and 'correctness' of a retrieved item." (op.cit.:260)

[1] Check on meaning of item
- [a] translate item into L1
- [b] translate cotext with or without item
- [c] compare concurrent items (i.e. alternative fillers)
- [d] incorporate meaning of co-text (including perceiving relationships with preceding or subsequent text)

[2] Check on form of item
- [a] structural analysis of item and/or co-text
- [b] translate item into L1
- [c] check by 'sound' of candidate filler item (repeated articulation of item or of alternative candidates)
- [d] recall of past situation
- [e] construction of example

[3] Other evaluation strategies which cannot be assigned to [1] or [2], or are seen as a mixture of [1] and [2].

[a] re-read all or part of co-text
[b] check by 'sound' of item (as above, but also including sounding' items embedded in co-text)
[c] check via other languages
[d] check via 'look' of written representation

[4] No identifiable strategy

As I have already suggested, above, this apparent clear distinction between 'recall' (i.e. recovery) versus 'evaluation' strategies may be open to question. How, for example, can we be sure that translation into the L1 is not as much a means of retrieving a word as of evaluating the recovery?

More empirically valuable, perhaps, is Feldmann & Stemmer's cline of 'bottom-up' to 'top-down' processing strategies–which can be seen as paralleling the 'local' versus 'global' spectrum posited by Trollope 1995 and others. Bottom-up C-Test strategies include structural or syntactic analysis as well as item repetition or 'sounding out', while towards the top-down end of the scale we find the repetition of longer stretches of text, inference of overall text meaning, and reference to extratextual information and prior knowledge. Behaviours such as these are, the writers tell us, directly reflected in verbal report data.

Lexical inferencing behaviour: Ames 1966

An early attempt at categorising contextual cues, and perhaps the first to make use of subject reports about the processing of cues, can be found in W. Ames' 1966 paper *The Development of a Classification Scheme of Contextual Aids*. This paper looked at native-speaker subjects' processing of a 'nonsense'-word lexical

inferencing task and, despite its vintage, it offers a more ambitious and comprehensive taxonomy of cues than do many more recent studies and merits looking at in some detail. Interestingly, Ames' classification is only indirectly concerned with the distance between a target item and cues to its meaning; far more central is the formal or semantic relationship of target and cue.

Ames' instrument

Ames derived target items ("contextual situations", in his terms) from twenty randomly-chosen general interest texts, hoping in this way to target a

> "more complete group of the possible contextual aids that could be used in determining the meanings of unknown words" (1966:61)

Confusingly, in considering the range from which target words might be 'randomly' selected Ames explicitly rejected a role for the content word–function word distinction. Instead he chooses to select from the "four main meaning classes" of adjectives, adverbs, nouns and verbs, thus ruling out "such words as determiners, auxiliary verbs...conjunctive adverbs, connectives [and] pronouns." Ames remarks here that words from the form classes selected contrast with

> "structure words, which are relatively empty of meaning themselves but which give coherence to a language [and] relate words and word groups."

A rose by any other name? In fairness to Ames, the content-/function-word dichotomy does not seem to have been widely used until after his paper appeared. Reasonably, Ames also excludes words such as proper nouns, dates, etc. on the grounds that these may be impossible to fill from contextual evidence alone.

Similarities with cloze procedure

Ames preparation of his target texts has several interesting parallels with cloze procedure. A fifty-word stretch of text is left untouched, and (provided it belongs to the class of candidate words) the fifty-first word is marked as a target item. Fifty more words are counted off, followed by a second target, and so on throughout the text. This procedure is analogous to a fixed-ratio cloze with a target-to-context ratio of 1.50.

The 334 words marked as target items are then replaced with a 'nonsense' simulated word (a form which I will henceforth refer to as a 'pseudoword') of approximately equal length to the original word. According to Ames, pseudowords are used in place of blanks "to more closely approximate a realistic reading situation" and are designed not to resemble or be open to confusion with authentic English words. Some semantic content is carried over, however, in that "structural or inflectional endings" are retained. Ames accepts that these endings, which potentially identify the item's word class at least, do offer cues to meaning, but argues that the retention of these endings is necessary in order to "keep the grammatical construction of the sentences intact". In other words, he does not appear to have seen the maintenance or preservation of syntactic structure as part of the reader-subject's task. The desire to preserve sentence structure perhaps underlines Ames' apparent conception of his task format as a measure of reading comprehension, rather than of overall language proficiency. In a test of the latter, there would be no reason *not* to require subjects to recover the grammatical class

of a target item as well as its meaning.

But our primary focus here must be on Ames' attempts to work up a classification of contextual cues. From a total of 334 contextual situations Ames derived fourteen types of cue. These are set out, verbatim, here:

1. Clues Derived from Language Experience or Familiar Expressions
2. Clues Utilising Modifying Phrases or Clauses
3. Clues Utilising Definition or Description
4. Clue Provided Through Words Connected or in Series
5. Comparison or Contrast Clues
6. Synonym Clues
7. Clues Provided by the Tone, Setting, and Mood of a Selection
8. Referral Clues
9. Association Clues
10. Clues Derived From the Main Idea and Supporting Details Pattern of Paragraph Organisation
11. Clues Provided Through the Question-and-Answer Pattern of Paragraph Organisation
12. Preposition Clues
13. Clues Utilising Non-Restrictive Clauses or Appositive Phrases
14. Clues Derived From Cause and Effect Pattern of Paragraph and Sentence Organisation

Despite its sometimes quaint formulation, the detail and apparent comprehensiveness of Ames' taxonomy calls for an effort to relate his scheme to those of more recent research. In each brief summary below, the figure in brackets shows the percentage of contextual situations in which that type of cue was identified, rounded to the nearest decimal point. All examples have been taken

from Ames 1966, and I have added my own reactions and comments to each category, rather than in a separate section.

Category 1 cues (7.8), above, involve the reader's recognition of a common phrase or idiomatic expression, such as: "written all over their *herats* (faces)" Clearly, we are dealing here with a kind of collocation or idiomatic cue.

Category 2 cues (9.3) involve the use of "phrases that [modify] unknown words or phrases which contain unknown words: "he...*shoered* (stabbed) her repeatedly with a knife".

Category 3 (6.6) cues describe or define the unknown word "somewhere in the text", as in: "...the professional *nodar* (donor). A few who sell blood..."

This particular category seems to deal with target–cue distance as well as a 'paraphrase' type relation.

Category 4 (9.2) cues are provided by "words connected or in series". Thus, Ames tells us: "the sonnets and *grods* (plays) of William Shakespeare"

"is an instance of two nouns connected by an 'and' with one of the two words connected unknown". All well and good, but what does this actually convey about how *grods* might be recovered? There is no indication in the subject response cited that (s)he found any *cotextual* cue to this item, so that the 'connectedness' of 'sonnets' and '*grods*' may well have less to do with the presence of 'and' (compare "..the sonnets, the *grods* of William Shakespeare..") than with lexical association

allied with prior knowledge .

Category 5 cues (11.1) depend on relationships of comparison and contrast: "...will it be a blessing or a *fome* (bane)?"

One response quoted shows that the subject, while unable to come up with a suitable replacement word, clearly realised that the word must involve some contrast with the idea of 'blessing'. Recovery of this item appears to rely on a knowledge of the logical relationship: X *or* Y.

Category 6 cues (10.8) are those provided by synonyms in the text: "they *telded* (wanted) piped-in music, but no 'specials' announced over the loudspeakers. They wanted the candy put out of the reach of children."

Here the synonym is supposedly a repetition of the target item in the following sentence, but it is hard to see how this would permit 'wanted' yet rule out, say, 'demanded' or 'required'. Any plausible alternative filler would, however, carry a semantic load expressible as [+ WANT] or [+ REQUEST] and Ames' criterion of acceptability requires no more than an approximate semantic equivalence of deletion and recovery.

Category 7 (8,7) cues of 'tone', setting', and 'mood' seems to be indistinguishable, i.e. the three labels appear to be alternative descriptors for the same class of cue. Cues available here include "...use of specialized terms or vocabulary", so appear to be tapping what we might nowadays label knowledge of passage topic or genre.

Ames (op.cit:74) appears to see this class of cue as mainly intratextual ("...the reader's understanding...through previous events and actions described by the writer...") but in the examples he provides, extratextual knowledge must surely be involved too. For example, an in-text reference to 'the space age' surely helps cue 'astronauts' for *obspedists* , just as 'trembling' must help to cue 'chill' for *maodly*.

Category 8 (3.9) cues are labelled "referral cues", and make use of "certain referral signal words" such as 'these', 'same' etc. All examples cited by Ames pertain to anaphoric relations. From the examples cited, this category appears to overlap with that of synonym cues (above): "Look at the figures for deaths... [New sentence] These *wharfabins* (statistics)..."

Category 9 ("association") cues depend (op.cit:75) on "certain words used by writers [which] aroused associations in the minds of the readers" These cues clearly depend on extratextual (in modern terms, 'schematic') information and divide into noun--verb association (to *play* the new symphonies) and adjective--noun associations (boys in short *pants*). These might also be seen as collocation-type cues.

Category 10 cues (5.1) depend on readers' perception of 'main idea--supporting details' paragraph organisation. as when an unknown word in the 'topic' sentence is substantiated in a following sentence, or vice versa.

In category 11 cues (2.7), the target word is either part of a question, from whose subsequent (in text) answer the meaning can be derived, or part of the answer so

that the meaning is derived from the question.

Category 12 cues (6.0) are labelled 'preposition' cues in Ames' scheme. In such cases, the target item is the object of a preposition, and the reader explicitly draws upon this relation in arriving at the target's meaning. The examples Ames cites are not entirely convincing in this respect. In the example below (op.cit:78), the subsequent noun phrase 'highway patrol officer' may be as salient a cue as the preposition itself:

> "A little later, as he sped northward along a California *cliotol*, Kendricks was stopped by a highway patrol officer."

Category 13 cues (2.7) come into play where the target is "in either a non-restrictive clause [or] an appositive phrase within a sentence" (op.cit:79) Here the subject makes use of the relationship of that clause or phrase to the rest of the sentence:

> "...after delivery a mother's stay is limited to 24 hours–*mantly* a sufficient period for her infant to 'stabilize'."

Yet here too prior knowledge must play a role. In the absence of topic knowledge of childbirth, how could a reader judge whether twenty-four hours is 'typically', 'invariably', or 'seldom' sufficient for a baby to stabilise? Of course, the required topic knowledge may be contained elsewhere in the text.

Finally, category 14 cues (3.0) made use of cause and effect relationships such as: "...by cheating the insurance company [drivers] are only pushing their own

premiums *lorter* (higher)." Cause and effect relations obviously require a general knowledge of the world

Twenty-five (7.5) "miscellaneous situations" were identified, in which the contextual cues used did not seem to fit into any of the categories above. Given this low percentage, Ames sees no value in creating further categories. As even lower percentage figures are cited for some of the categories above, however, we can infer that these miscellaneous situations displayed too little regularity to warrant the creation of new categories.

Interestingly, in a large proportion of the responses Ames cites, the subject replaces the pseudoword with the exact original word. This might appear suspicious, but it may be that Ames selected for citation only exact-word responses from those he collected–perhaps hoping thus to show that 'exact' cued recovery is possible.

Rather more remarkably, the occurrence figures for the cue categories above neatly add up to 100% of all cue exploitation identified. This means that either (1) in no instances were subjects observed to make use of cues from more than one category, or (2) only one 'cue-uptake' was recorded for a given target. The first of these would be surprising in the light of the finding (Jonz 1991) that respondents commonly identify more than one contextual feature cueing recovery of a target item.

Nor is it uncommon for respondents to be uncertain as to the relative salience of cues in triggering recovery of meaning: recovery often appears to be an incremental process by which possible fillers are initially activated by a given cue or cues, and their suitability confirmed or disconfirmed by other cues again. This must cast serious doubt on the notion that a single cue can reliably be allocated the determining role. Ames uses the phrase (:76-77) a "main aid in determining the [meaning]", but gives no indication of how any distinction between main and subsidiary aids was reached.

What insights can we glean from Ames' ambitious but often seemingly rather arbitrary scheme of cue classification, with its at times dated terminology (Who nowadays talks simply of 'association' between words, and can we interpret the term in the way a writer of almost forty years ago may have meant it?) One fairly obvious lesson is that it is not easy to isolate more than a limited number of target–cue relationships.

Haastrup 1991 and lexical inferencing

Lexical inferencing tasks in which the target items genuinely lack morphological cues to meaning might be seen as more or less analogous to cloze deletions. Studies of lexical inferencing such as Haastrup 1991 frequently focus on such (real- or pseudo-)words, which by virtue of their lack of 'internal' cues force the comprehender to rely on contextual constraint (inclusing extratextual knowledge.) For our purposes here, such items may be seen as functionally equivalents to cloze

blanks. Haastrup 1991 offers the word 'squalor' as an instance of a word effectively lacking in morphological cues to meaning. Apart, possibly, from the Latin nominal marker *-or*, there does indeed appear to be nothing in the word itself which might cue meaning for Danish L1 subjects, or indeed for English L1 subjects innocent of etymology:

> "Paper everywhere, piles of it. And books too, in hardback mostly. Squalor, yes, but *academic* squalor."

Haastrup's taxonomy of cues

Haastrup 1991 lists three categories of "knowledge source":

CONTEXTUAL CUES
- the cotext [i.e. the task passage]
- knowledge of the world

INTRALINGUAL CUES
- the test [target] word
- the syntax of the sentence

INTERLINGUAL CUES
- the L1
- Ln [i.e. any language except L1, TL]

Haastrup's category 'knowledge of the world' is fairly simply defined (1991:94) as "what the informant...cannot have taken exclusively from the text". Knowledge of the world, then, underlies what Haastrup labels 'metacognitive' statements such as "That sounds reasonable", as well as explicit references to extratextual knowledge. Of Haastrup's 'intralingual' cues, the first ('the test word') is clearly

not relevant to cloze processing. Her second sub-category ("the syntax of the sentence") is glossed thus: "[T]he informant uses the structure of the sentence in which the test word occurs." It is not immediately obvious why this category (which Haastrup does not further elucidate) could not have been included under the heading of 'cotext', although we may reasonably infer that she wishes to distinguish between the subject's reference to sentence-level syntax from that to informational/lexical information in the passage as a whole.

Haastrup's heading of 'interlingual cues' encompasses phonological and orthographic similarities between the target word and an L1 word. Thanks to the physical absence of a target word in cloze (and assuming that all blanks are physically indistinguishable) these factors can play no role in cloze process except perhaps in cases where that an L1-derived item is used to fill a blank in the belief that it is an appropriate TL word. The role of non-target languages, in the main the L1, appears to vary considerably in lexical inferencing tasks (Haastrup 1991; Aizawa 1998). Their role in cloze recovery is also far from straightforward. As mentioned above, it does occasionally happen that a test-taker will enter an L1 item (or even more occasionally an L3 word) in the belief that it fits the blank. Clearly, a non-English word cannot be even SEMAC correct in a cloze, but the choice may give some indication of how well the test-taker has understood the context. If cloze were to be used as a test of overall reading skill there might be no reason to insist on TL fillers!

Beyond these misconceptions, however, it is clear that the informant's own L1 can and often does play a role in her overall processing of the cloze passage. This is taken up in the following section, but it is worth pointing out here that while not all of the taxonomies set out in Figure 3.2 have an explicit place for translation of one kind or another, this aspect of processing may in theory at least be *redundant* in verbal report tasks such as think aloud–that is, the cognitive and/or linguistic demands of the task may be such that informants do not need to translate; it may be *suppressed* in that informants are deliberately not translating, for whatever reason; or it may be *silenced*, in that informants are doing it but deliberately or otherwise leave no overt record of translation. Either of the latter two situations may be due to an explicit or implicit instruction to report only in the TL, to informants' internal perceptions of the researcher's ability to understand their L1 (or dialect thereof), or to their conception of 'good' language learner behaviour or other notions inculcated in the course of their formal language training. Post-task discussion with informants has elicited claims for all three states, above, but retrospection alone–unsupported by concurrent data–is a weak basis on which to accept them,

3.14 Interlingual cues at the supra-word level

So how do informants' L1s or L3s (Haastrup's L*n*) impact on cloze processing? While no subject can see in a cloze passage blank a resemblance to a *word* in her own L1, interlanguage parallels at phrase or clause level may play a role in recovery. For each of the examples below, at least one informant explicitly

identified a parallel L1 construction as a cue to recovery:

> Cloze target sequence: '[...] It gets on my nerves [...]'
> Parallel sequence informant's German L1: *Es geht mir auf die Nerven*
> Cloze target sequence: '[...] Shall we go out together?[...]'
> Parallel sequence informant's German L1: *Wollen wir mal zusammen (r)ausgehen?*
> Cloze target sequence: '[...] They were known to] drop dead from strain [...]'
> Parallel sequence informant's German L1: *[...] die fielen vor Überanstrengung tot um*)

The extract below from a think-aloud protocol shows clearly that spontaneous comparison of L1 and TL collocational or structural parallels occurs in cloze processing, although this may lead to error as often as to an acceptable recovery. In this instance the informant immigrated in late childhood from Croatia to southern Germany but was apparently at native-speaker level. The immediate context of the deletion (from the text of a letter) was 'I was relieved to hear that you were both safe and well':

"[...] I was.. to hear that you were both safe and well.. I was [INAUDIBLE] to hear that you that you both.. are well.. are.. both.. healthy?..ARE..BOTH..YOU ARE BOTH WELL? [...] AH I FELT RELIEVED (erleichtert) WHEN ... I was *lighter*? [LAUGHS] NO.. I was [6 secs] DON'T KNOW WHAT I SHOULD PUT IN HERE [...]"

This behaviour is unsurprising; there are enough syntactic and idiomatic overlaps and similarities between English and German for phrase-level or idiomatic L1 counterparts to be potentially valuable as cues to meaning. Many informants

appear to be conscious of this feature and its potential utility in language processing, although we cannot assume that every application is deliberate. A number of my think-aloud informants have denied (in post-task interviews) that they made use of translation while completing cloze tasks, and this even though their protocols contained sometimes quite extensive paraphrasing or even close translation into the L1. While in one or two cases I suspected that these denials might be connected to self-image, other informants' apparently unfeigned surprise at hearing themselves render passage content into their L1 left me to wonder whether this behaviour was not somehow highly transient, or even barely conscious. I will have more to say about translation behaviour in my discussions of think-aloud and other data in chapters 5, 6 and 7.

3.15 The nature and function of 'guessing'

The taxonomies of cloze processing behaviour by Maghubai 1990 and Trollope 1995 both employ the category 'guessing'. Two criteria seem to be available in bringing this category into play: (1) the informant explicitly claims to be guessing or (2) the researcher makes the assumption that the informant was guessing. (Note that some writers have a category like 'other', where researcher assumptions of guessing may also end up.) I have already indicated that I am skeptical that truly random guessing, in which an informant pulls a word out of thin air, as it were, plays much role in cloze. Obviously, however, the informant's basis for choice must impinge on her consciousness strongly enoughto be noted (even if not reported under task conditions) otherwise it may appear to her that she has indeed

randomly picked the word.

Something like guessing in the sense of selecting among a range of possibilities may, however, play a part in cloze recovery. In the extract below, the informant gives no indication of how she selected among the various types of sport event (swimming; jumping: horse-racing; boxing) she had listed as possible fillers for OLYMPICS deletion (21):

"[...] I don't know which.. have to choose though.. well I'll take swimming.. probably it's wrong [...]"
(JL1 think aloud Masako)

although she does claim to lack confidence in her choice. When an informant explicitly claims to have 'guessed' a cloze filler (or word meaning, in lexical inference) she may be saying as much about how much confidence she has in her selection as about how she arrived at it. I set an intact group of second year Japanese university undergraduates (n=22) a 10-item rational cloze/gap-filling task along with a brief checklist of cue-locations-cum-processing behaviours such as 'use of passage content outside the paragraph containing the deletion', 'use of extratextual knowledge', etc. The final item on this list was the open category 'Other.' Respondents were to identify with a 'tick' the operations which had played a role in their filling of each blank, and if 'Other' was one of these they were to write (in their L1 or in the TL) what the 'other' behaviour was.

All respondents received this task sheet, but half the sheets also carried a six value Likert-type scale beside each item, along with the additional instruction that respondents should indicate on that scale their confidence in the correctness of each recovery. The null hypothesis was that the presence of absence of a scale was not related to the frequency with which respondents would claim (under the category 'Other') to have 'guessed' the fillers. Following the session, 'Other' category entries interpreted by a native-speaker of Japanese and myself to mean 'guessing' were totaled on both kinds of task sheet, and a chi-square test carried out using the software package SSP v.2.0 (which includes Yates correction for 2-by-2 tables) This provided a chi-square value (4.4370 at *df*1, significant at .05) which suggests that the availability of a scale via which the informants can indicate confidence, or lack of same, in a filler may be related to the frequency with which they claim to have guessed it. The claim to have 'guessed' a filler, then, may in some cases be in large part a means of claiming (with whatever motive) low confidence in it. This inference fits the incidental finding that in 69% of those instances in which informants whose task sheets carried the 1–6 scale also claimed to have guessed a filler, the associated scale value was 2 or lower.

3.16 Cross-referencing of taxonomies

To round off the discussion of classifications of types of processing behaviour, strategy, or what you will (although the connection between these and the upcoming taxonomies of target-cue distance will be clear), Figures 3.2 and 3.3 attempt to chart my interpretations of how the taxonomies of processing

behaviours discussed here cross-compare with one another. As the number of items in parentheses suggests, there was a good deal of uncertainty about the task of cross-matching, and the table represents no more than a tentative attempt. (The table appears on the following page.)

Recovery behaviours

	Mangubhai 1990	Block 1986/92	Feldmann & Stemmer	Trollope 1995	Allan 1992	Ames 1996
Local context syntactic	Task A, B		Recall 1a	25(?)	Recovery 1?	12(?), 13
Local context collocational/lexical	Cognitive D (C?)	15(?)	Recall 3b(?)		Recovery 1?, Recovery 4(?)	2, 4, 5(?), 12(?)
Global context	Task C, D			6, 16, 15(?)	Recovery 2(?), Recovery 6	5, 8, 10, 11, 13(?) 14
Retrieve item class only	Cognitive E					
Extratextual knowledge	Cognitive A	1, 3(?), 6		18, 1(?)		7(?), 12(?), 13(?)
Explicitly fill provisionally				28		
Paraphrase	Cognitive G	11		21		
'Logical inference'		5		8(?)		
Translate			Recall 5a–c,	22, 29, 30		
Read aloud/ repeat	Cognitive C			26	Recovery 4	
Query meaning of extant content		13, 14		23, 24		
Affective response to content		4		12, 14		
'Guess'	Cognitive F			20		
Skip item				7(?)	Recovery 5	
Abandon item	Task G, H			19		
'Metacomment'				5, 11, 10(?), 28(?)		
'Automatic' recovery	Automatic A, B					

Figure 3.2: a tentative cross-referencing of six taxonomies of processing behaviour relevant to cloze

Checking / revision behaviours

	Mangubhai 1990	Block 1986/92	Feldmann & Stemmer	Trollope 1995	Allan 1992	Ames 1996
By 'sound'	Judgement A		Evaluation 2c, 3b(?)	4?		
				4?		
By 'look'				4?		
Via L1 glossing			Evaluation 1a, 1b, 2b	4?		
Via knowledge of syntax rules	Judgement B		Evaluation 2a	4?		
By reference to passage content	Judgement C			4?		
Via extratextual knowledge	Judgement C(?)		Evaluation 2d	4?		

Figure 3.3: a tentative cross-referencing of six taxonomies of processing behaviour relevant to cloze(cont.)

3.17 Another criterion: Target-cue distance

Moving away from attempts to classify the 'types' of meaning-recovery process which are set in motion as test-takers set about a cloze task, it is conventional to structure the cues they will potentially utilize in terms of target-cue distance, or the physical distance between the target deletion and the information which might or might not enable the test-taker to fill it.

It would be reasonable to ask what we mean by target-cue distance. The immediately obvious sense of physical distance is generally that of linear distance, readily measured by counting the number of words (or, if precision is vital, individual characters and spaces) between the target blank and the cue. This is the normal mode of expressing textual distance, but we should not forget (Taft 1991)

that reading is not *entirely* a linear process: information may be taken in through parafoveal vision from areas of the passage vertically below or above the zone of focus, as well as from those to the left and right. This information may or may not be heeded, or if heeded may play a role in the process of interpretation. The physical interval between the first two occurrences in this paragraph of the word 'distance', for example, is either six words or approximately 6mm. In an age of on-screen reformatting, kerning and font collections, the impracticality of measuring textual distance other than by word-count will be clear, but 'parafoveal heeding', as it were, is a real and unpredictable phenomenon in text processing.

The concern with the amount of text occurring between target and cue(s) stems from the question of the question of the level(s) at which cloze procedure taps subjects' processing of the passage. Identification of the cues test-takers use in recovering deleted items can provide insight here: if cues relatively distant from the target are being activated, then clearly cloze must be tapping something more than processing at a local level. Almost no-one, of course, doubts that relatively distant cues are at least sometimes tapped by natural cloze; the question is rather that of how often this occurs.

Schemes of target–cue distance

Schemes of 'linear' target-cue distance has come and gone over the years, with many of the survivors re-labelling, revising, or at least owing a clear debt to

Bachman 1985. Given its obvious influence, I begin with that scheme.

Bachman's 1985 (a rational cloze study which, surprisingly perhaps, goes unmentioned in Markham's 1987 review of the literature) offers a four-part classification:

(1) within clause
(2) across clause, within sentence
(3) across sentences, within text
(4) extratextual

Where short task passages are used this simple scheme may be quite adequate.

Sasaki 1993, in a study of the amount of text high and low (linguistic) proficiency test-takers required to fill fixed-ratio deletion cloze blanks refines Bachman's scheme. Taken over from Bachman 1985 are the 'within clause' and 'across clause, within sentence' categories, which together encompass what we can call 'immediate' and 'local' constraint. Sasaki notes, however, that Bachman's 'across sentence' category does not take account of paragraphing; she thus subdivides this category into searches 'across sentence, within paragraph' and 'across paragraph, within text'.

To this Sasaki adds an 'extratextual' category, i.e. information which is not contained within the text, but is assumed to be held within the subject's prior knowledge. The two final categories in Sasaki's 1993 scheme are (1) 'guessing' ("..because the examinees sometimes randomly chose answers..") and (2) 'missing' ("..or did/could not say anything about their test taking processes") The

resulting scheme of 'distance' relations is then:

(1) within clause
(2) across clause, within sentence
(3) across sentence, within paragraph
(4) across paragraph, within text
(5) extratextual
(6) guessing
(7) missing

To what extent does Sasaki's more elaborated taxonomy represent an advance over that of Bachman? Where paragraphing in task passages is both present and meaningful (I refer here to 'genuine' paragraph divisions that indicate a shift of topic, rather than the more or less sentence-based paragraphing found in newspapers and increasingly in magazines) the extra category may be useful in quantifying passage distance. A further criterion for meaningfulness however, might be that task-takers be aware of the significance of paragraphing as a rhetorical device. (something by no means inevitably true of second language readers of English.) Without this awareness, task-takers may be searching more widely, but not qualitatively *differently*.

Sasaki's latter two categories are behavioural rather than measures of textual distance. I find it hard to believe that test-takers truly guess cloze fillers, except perhaps in multiple-choice cloze. If an answer *was* arrived at randomly, it seems reasonable to suppose that the subject was unable to recover it by any other means. Verbal report data (not reported here) from students deliberately given too little

time to complete a gap-filling task suggests that even when rushing to finish they typically do not resort to random guesswork. If random recovery of items actually exists, it may be an indication of very serious processing difficulty. Sasaki's final category of 'missing' is a kind of necessary evil. It represents a failure on the part of the informant to reveal the textual cues she used, or the researchers failure to perceive any revelation that occurred.

I have already discussed Manghubai's 1990 two-part scheme, according to which comprehenders look either at the 'immediate context' before generating a word or at the 'larger context'. Another taxonomy of distance that goes in the other direction to Sasaki and Haastrup by postulating fewer categories than Bachman 1985 is Allan 1992 who offers a three-part classification of cue-seeking behaviour:

(1) search local context around gap
(2) search text up to but not beyond gap
(3) search text beyond gap as well as text before, (i.e. the text in full or in any part)

Although existing taxonomies may not transfer well or at all to a different data-gathering context, they are still open to evaluation on the basis of another researcher's own observations. On the basis of my own informants' protocols I would recast Allan's "up to but not beyond [the] gap" as something like 'up to the gap and within its local context to the right': except where the missing item identifiably comes at the close of a sentence or clause, my experience has been

that verbal reporters will typically continue to read aloud (or claim that they read silently) to the end of the unit they perceive as enfolding the (in cloze or gap-fill) missing item. With that minor change, and the rider that what follows is true of most schemes, Allan's taxonomy fits with the way some of my informants have perceived their own cue-seeking behaviours.

Haastrup 1991 also allows a role for extratextual information, which she distinguishes from that available in the co-text' (i.e. the passage itself.) This category she further differentiates (op.cit.:93ff) into:

(1) "one or two words from the immediate co-text"
(2) "the immediate co-text"
(3) "a specific part of the co-text beyond the sentence of the [target] word"
(4) "unspecified use of the co-text"

'Immediate co-text', for Haastrup, clearly means the sentence in which the target item is located; this is large enough a unit to make it more precise location of cues desirable at times, and 1. appears to cover situations in which the cue is part of the same syntactic structure, phrase or collocation as the target: two to three words *on either side* of the blank will embrace most such 'chunks'. Haastrup's third category may appear rather loosely defined, but it should be kept in mind that her stimulus passages ("co-texts") were sometimes very short.

As for category (4), it seems implausible that text comprehenders rely on an undifferentiated 'global' context, rather than on specific elements within it. The problem, of course, is that these elements cannot always be identified from

informant protocols or other data, so that there is indeed a case for setting up a category of unspecified cotextual reference much like Sasaki's "missing". Haastrup's criterion for extratextual information is that it must provide cues which could not possibly have been derived from the co-text itself. In my own data-gathering I have not always been able to confidently draw this distinction between inference from co-text and that from prior knowledge. Given that Haastrup's concern is less with the precise location of cotextual cues than with overall patterns and weightings of cue activation, her scheme appears to be adequate.

Haastrup's categorisation of target–cue distance is slightly more precise than Manghubai's 1990 simple dichotomy, according to which comprehenders look either at the 'immediate context' before generating a word or at the 'larger context'. Two plausible reasons for this are that (1) even 'rich' verbal report protocols may fail to indicate precisely where in the co-text the informant is taking up cues and–more importantly–(2) Haastrup's concern is less with the precise location of cotextual cues than with overall patterns and weightings of cue activation. Again, I round off this discussion with a tentative cross-tabulation of schemes of target-cue difference by a number of writers

Bachman 1985	Markham 1987	Maghubai 1990	Kesar 1990	Sasaki 1991	*Haastrup 1991	Allan 1992
(1) Within clause		(1) immediate context	(1) word level / part of sentence	(1) within clause	(1) "one or two words from the immediate co-text"	(1) search local context around gap
(2) Across clause, within sentence	(1) intrasentential		(2) sentence level	(2) Across clause, within sentence	(2) "the immediate co-text"	
			(3) intersentential	(3) across sentence, within paragraph		
(3) Across sentences	(2) intersentential	(2) larger context	(4) whole text level	(4) across paragraph, within text	(3) a specific part of the co-text beyond the sentence of the [target] word	(2)?; (3) search text beyond gap as well as text before
(4) Extra textual	(3) pragmatic		(5) extratextual	(5) extra textual	(5)	
			(6) metacognitive	(6) guessing		
			(7) other	(7) missing		

Figure 3.4: A tentative cross-referencing of seven taxonomies of target-cue distance in studies of cloze (and *lexical inferencing) tasks

3.18 Categories with low explanatory value

Taxonomies of strategies often involve categories with apparently limited explanatory value like Sasaki's "missing" or Haastrup's "unspecified use of the co-text." I would argue that such a category is desirable or even necessary for an

honest accounting of how much data a given reporting procedure elicited, and how much of that data could in fact be interpreted.

Where a 'missing' category represents absent data it is important to remember that this occurred either because the subject *could* not or *did* not verbalise (I have yet to encounter a case of deliberately *would* not do so.) about a target item. In the former case, and we can seldom be entirely sure which it was, we might hypothesise an 'automatic' recovery, which in turn suggests a lack of conscious processing. If the subject *was* conscious of processing that she did not verbalise, this may have been due to a very high processing load which left very spare cognitive capacity spare for the task of reporting.

It may, however, have stemmed from other factors: simple distraction from the verbal report task (as in the case of an LL informant whose sustained on-task silence was due, it turned out, to a deep interest in the workings of her booth's console); a belief that there was nothing that merited reporting; or even (I have encountered several such instances in post-task interviewing) to the realization that her TL resources would not allow her to report a given inference in that language–her *self-selected* language of reporting. The reasons why data can go missing are so various that it is risky to read meaning into the fact that it did so.

3.19 Extratextual constraint an equally-available resource?

The question touched on above, namely how to conceptualise extratextual clues to meaning, merits some attention. Extratextual information might on the one hand

be seen simply as constraint at a greater 'distance', in terms of psychological 'availability', being outside the particular text currently under processing, but as think aloud protocol data reveals (chapter 6) it appears in some instances to be almost instantaneously applied. But is extratextual information available to all task-processors? On one level, obviously not, for we can reasonably assume that for any given topic there are people who know more, or less, about it. Equally clearly, subjects processing a cloze task whose topic is, say, the Olympic Games are likely to make use of information gleaned from other sources they have encountered. Native-speaker consultants from both German and Japanese educational backgrounds have assured me that 'their' education systems do offer a certain level of knowledge about this topic, so that all of my informants could be expected to know something about it. Based on my own experience of school, this may err on the side of optimism, but there are obvious difficulties in trying to assess respondents' existing knowledge levels prior to exposure to the task.

Actual possession of knowledge does not invariably lead to its application in practice, and in chapter 4 I report the findings of a (very) small study in which I tried to investigate the suggestion by several Japanese educators that Japanese L1 students and their German L1 counterparts might approach cloze with differing expectations about the role of extratextual knowledge in item recovery., such that Japanese informants might be less likely than their German peers to apply whatever prior knowledge of the topic they had.

3.20 Conclusion

This chapter has attempted to review some of the taxonomies of constraint that are presumed to operate in cloze, and at some categorizations of cloze test-takers' processing behaviours. It has been noted that researchers into the nature of cloze test-taking strategies have employed some of the categories of researchers into reading of continuous text and of tasks such as lexical inferencing, and a reasonable inference would be that however much these tasks in fact have in common with cloze, the intuited overlap is considerable.

While I would agree that reading, cloze and other tasks do all have something in common beyond the mere fact of textual content, I admit to a certain scepticism about the label of 'strategy', a term whose meaning has (not, as we saw in chapter 2, unlike that of cloze itself) come to be applied to a wider range of events than seems appropriate. I will therefore do my best to substitute the more neutral labels of (processing) 'event', 'operation' or 'behaviour'. These may be taken to cover both deliberate and non-deliberate occurrences during task-processing and the attendant reporting.

In the following chapter I describe how I went about setting up the think aloud study whose data is reported and discussed in later chapters and look at some research into verbal reporting as a data-elicitation procedure. As will become clear, I have some criticisms of what is often seen as the standard model, that proposed by Ericsson & Simon 1984/1993, and I will try to illustrate why I think the model is perhaps unrealistically restrictive for use with language processing tasks such as

cloze. I ought to moderate this suggestion by noting that I see the revised (1993) edition of Ericsson & Simon's work as having subtly shifted the ground in verbal report theory, and I suspect that I have not yet reached a proper understanding of all of the implications of this apparent move towards a more liberal model. As Boren & Ramey 2000 note, however, the greatest impact of Ericsson & Simon's work was created by their original 1984 publication, and it is on this which (Boren & Ramey 2000) many researchers still appear to base their arguments.

CHAPTER 4: SETTING UP THE THINK ALOUD STUDY

4.0 Introduction: 'self-as-subject' explorations

The paradigms of experimental science, or at least its terminology have until comparatively recently been dominant in the linguistic sciences, with the resulting perception of ethnographic data as essentially 'anecdotal' or even (dread word) 'subjective'. Today, however, there is a growing body of research which carries ethnographic approaches to the lengths of allowing the researcher himself or herself the role of the described as well as the describer. Cohen 1984 is one a number of writers who discuss the use of diaries (or other forms of self-report in which the researchers record their own experiences of language-learning or linguistic interaction) and who argue that these offer new and 'thicker' insights into the processes of language acquisition and language use.

'Participant observation' (Cohen, Manion & Morrison 1996:110ff) in which the researcher actively takes part in the events or activities under study, has long been a staple of ethnographic research, and has found its way into the repertoire of researchers into language acquisition and use. These research formats can, like verbal report, provide data not readily recoverable by other methods, and few would wish to return to the days when (as a veteran researcher in child psychology tole me) if it couldn't go on a graph, no one took it seriously. Keen to experience for myself the kinds of task I might be asking informants to complete, to find out how readily I could interpret my own problem-solving behaviours let alone those of others, and of course to note any missteps that such an exploration might help me avoid, I resolved to use myself as my own think-aloud-about-cloze informant.

4.1 Acquiring the self-as-subject passages

I had a helper select at random from the pages of a reasonably sophisticated German news magazine ("*Der Spiegel*") a number of articles in the 500-1000 word range. She then mutilated the articles according to standard cloze criteria (the introductory sentence or two left entire, and subsequently every seventh word deleted) and photocopied the result. This produced a set of cloze tasks authentic in all respects save that the physical size of blanks varied according to the word deleted—thus providing a clue as to word-length. This had to be accepted: I did not feel able to ask an unpaid helper to retype the passages with blanks of equal length, and had no access at the time to adequate scanning and OCR facilities.

Contrary to my expectations, in actual processing of the passage I became conscious of the usefulness of the deletion-length clue itself only in insofar as it helped me to *reject* specific candidate fillers which seemed too long for the space: I recall no instance in which awareness of 'space available' helped me to come up with candidate fillers, or even clearly delimited the class of word that might belong. One possible exception to this, gleaned from a think-aloud audio-recording I made at the time, is that the length of a blank seems to have allowed me to intuit, as soon as it came into foveal vision, whether a deletion might contain an adjective (*mehrere Kinder*; several children) or could only contain an article or other short determiner (*die Kinder*; the children) The notorious length of many German composite nouns does not seem to have played much of a role in this regard.

In the course of a discussion on this point, a colleague in Berlin remarked that he had (in his own word, naïvely) for years simply whited-out words in texts to produce cloze passages for didactic use in his reading courses. More recently, he had begun to use computer print-out versions of the same passages as produced by cloze generating-software. He had not, he claimed, noted the subsequent fall-off in average scores that might be expected if deletion-length indeed played a significant role in cloze recovery. From my processing of three German cloze passages on a range of topics (FLICK, focusing on a 1980s financial scandal in Germany; AETNA, concerning a threatened volcanic eruption, and KINDER, focusing on violent childhood behaviour) I arrived at a tentative list of behaviours or operations I felt I had employed in filling blanks, (and, importantly, in comprehending the remaining, extant passage text.) This is discussed below.

4.2 Developing a classificatory scheme for think aloud

The route outlined above is just one of at least four possible approaches to the setting up a taxonomy of task-processing events or operations with which to analyse think aloud or other verbal report data:

1. The researcher may adopt an *a priori* scheme in whole or in part (with perhaps some emendations) from previous research.
2. She may set up an *a priori* scheme based on those processes she has simply intuited to be relevant to task completion, prior to the gathering of any data.
3. Alternatively, a data-driven system may be used in which the researcher 'naïvely' collects data from informants and then attempts to isolate regularities or patterns of processing it may contain.

4 Finally, the researcher-as-informant may, perhaps as a first step, self-report on the processing operations she carried out.

The last of these is the option I initially employed, and I would argue that the resulting set of behaviours or operations is to some degree privileged by my deeper insight into my own processing. Privileged introspection is still introspection, however, and any scheme of categories developed on that basis must be tested against the data of other informants.

Each of the approaches outlined above has positive aspects as well as drawbacks. The first has the advantage that it allows for far greater comparability of findings across studies—a much undervalued (Faerch & Kasper 1987) aspect of research. This approach, however, carries the risk that the adopted analytical scheme may be inappropriate or inadequate to the data collected. From my own survey (cf. also Pressley & Afflerbach 1995) studies which comprehensively adopt the taxonomy of another are rare, although lip-service is commonly paid to the desirability of doing so.

The second option carries with it the danger that a researcher will structure the data according to an intuitive scheme at odds with potentially important features of the data. Pressley & Afflerbach 1995 survey reveals, I think, how widely classificatory schemes of ostensibly related data (i.e. all derived from studies of L1 reading) may diverge.

The fact that my own classification process began with the third and fourth options outlined above ('a scheme in which the informant self-reports on the processing operations she has carried out' and 'a data-driven system in which the researcher collects data and then attempts to isolate regularities or patterns of processing it may contain') derived from the fact that I had available data gleaned from trial (mainly, but not only self-as-subject) think aloud sessions, but also from a principled decision to avoid intuitive 'armchair' categories not based on concrete data. At this stage, however, I had encountered only one or two short studies which had made use of think aloud procedures in the investigation of text-processing. My initial categorization was thus almost entirely naïve, though derived from actual think-aloud data. Having made this first step, it was possible for me to refine and extend the scheme on the basis of further data. As I came upon investigations which paralleled my own, I attempted to compare my classificatory scheme with those of other researchers.

This latter step was made rather more difficult by the fact that many published reports fail to illustrate or exemplify isolated processing events clearly enough or often enough to allow genuine comparisons to be drawn across studies. Intuitively, however, researcher A's 'processing event X' often seems to be more or less equivalent to researcher B's 'processing event Y.' I thus felt able to map certain operations I had derived from my own trial data collection onto those described in other reports. Although these mappings could only be tentative, they made it possible to compare taxonomies expressed in sometimes far-from-identical terms.

Some of the more useful studies are outlined in the literature review chapters, but it must be borne in mind that studies of cloze processing using verbal reports (i.e. think aloud, retrospection, etc.) are still comparatively few. It is worth asking in this regard whether a high level of cross-study comparability may realistically be achievable (and indeed an appropriate goal) in situations where researchers are looking at different sets of processing operations, or the same processing operations across markedly differing populations or tasks. In refining my own set of processing operations I have been informed by studies which belong in one or more of the following categories:

(1) studies which incorporate concurrent verbal reports of any type of text-based task, e.g. Haastrup 1991
(2) studies which look at inferencing (and in particular inferencing during text processing) e.g. Trabasso & Suh 1993
(3) studies of the processing behaviours of first and second-language readers, where some kind of introspective data was collected, either through a VR procedure or by post-task interview, e.g. Cohen 1984
(4) studies of the mental operations involved in translating from or to a second language, e.g. Ridley 1997, Krings 1988

The most comprehensive categorization of mental events involved in text-processing is probably that contained in Pressley & Afflerbach 1995. This volume presents a compilation of 'strategies' drawn from a wide survey of VR studies of reading, but omits to illustrate this taxonomy in the analysis of readers' processing of extended passages. My data-driven work-up of a classification system had already begun by the time I encountered this—and many other sources—and I was reluctant to abandon my own categories in favour of a

second-hand taxonomy relating to an arguably overlapping by no means identical set of task requirements. It was, however, reassuring to find that some the behaviours I had identified in the data gathered to date appeared to be echoed in the findings of other researchers.

4.3 The self-reported operations of others

Hoping to gain further background insight, as it were, into the kinds of behaviours I might expect my informants to display, I followed up my self-as-subject explorations with a small investigation in which I asked two native-speaker teachers of English at a Berlin-Brandenburg *Volkshochschule* (a publicly-subsidised evening school for adults) and four (German L1) non-native teachers of English at an 'academic' secondary school (*Gymnasium*) in Berlin to work on a short narrative passage (FUNCHAL, see below) offered to me by another German secondary school English teacher who claimed to have used it successfully in a semester exam. I asked these consultants to work (in solo condition, in my presence) on the cloze passage produced from the FUNCHAL text, and to tell me in real time about any behaviours or steps which they found themselves using. They were also asked to comment on which (if any) of these behaviours they would expect *Abitur*-level ('the German A-Level') or undergraduate informants to employ in completing a similar passage. My intention here, of course, was to assess the validity of my own taxonomy-in-progress of task-processing behaviours and events. Some of these conversations were, if the consultant agreed, 'interim-recorded' (A number of my German informants were willing to be audio-taped while thinking aloud only if

the tape would be transcribed and the erased within an agreed period; other confidentiality conditions applied.) for later reference:

To aid the task of labelling anticipated behaviours, and thus to provide a maximum of comparability, I sat beside each consultant as she made up her list. I did my best to restrict my input to offering labels for behaviours, however, and tried not to appear to validate or invalidate suggested categories—even where a consultant clearly wanted my approval. I typically tried to give consultants the impression that I was hearing their ideas for the first time (which was often true) found them interesting and potentially helpful (ditto) but naturally could not evaluate them without further thought. One brief transcribed and translated exchange is shown below. The knowledge source-cum-behaviour this consultant was proposing was in fact later found to be indistinguishable in practice from 'knowledge of the world' and so abandoned:

Researcher

"So, if I've understood you correctly, you think that [students] might be aware of [similar texts] that [students] had already read in their own language?"

NNS T (R)

"Yes. I think one of our English textbooks has a text like this, about a [19th c.] traveler."

Researcher

"Can you think of a name [...] to give to this kind of behaviour, if students do it?"

NNS T (R)

"Perhaps 'think of [...] similar things in German?'"

Researcher

"Uh-huh. How about 'Think of a similar text in the first language?' Or 'similar stories'?"

NNS T (R)

"Yes, 'similar stories' is better, I think. In the end it's stories that people remember."

The diversity I noted among the resulting lists of behaviours proposed by informants (see below) suggests that this procedure may be of limited value unless a sufficiently large sample of these intuitive taxonomies is drawn upon to provide a clear idea of how widely an individual's suggestions are shared.

Despite my efforts to encourage consultants to offer concrete behaviours linked to specific cloze items, this did not occur spontaneously in every case. If gently pushed to link an operation to a specific deletion, however, they were usually able to do so within a few seconds. What this suggested to me at the time was that individuals were listing behaviours or operations they intuited as *potentially* usable in those contexts as well as those they had actually employed. I found it impossible, however, to draw a clear line between what I took to be intuited behaviours and those that had actually been applied. I doubt that my consultants would invariably have been able to make this distinction themselves, in fact, for experience (in my self-as-subject trials) cannot be unique: in some instances, one may become aware, at almost the same moment, of more than one route to recovering a cloze item or inferring the meaning of an extant passage item. I found it impossible, in such cases, to know which route I had followed, or whether both had played a part.

Also interesting is that only English native-speaker consultants explicitly drew attention to 'intuition' as a strategy, although the concept and the (morphologically

almost identical) equivalent term are at least as widely used in German as in English) and that only two consultants explicitly mentioned (though not necessarily in these words) reading ahead beyond the sentence, a strategy which is quite widely cited (cf. Oller & Jonz 1994) as useful in cloze test-taking. Two other consultants did however refer to the pre-reading of the passage, which they may have intended to embrace 'looking ahead'. (Recall that pre-reading of the entire passage is, according to studies cited in Cohen 1984, a strategy employed by only around 25% of cloze test-takers.)

The consultants utilise d above had been unable to devote much time to helping me, and so it was difficult to ask them to do more than intuit behaviours they felt informants might use. As a follow up to the above, I trialled the idea of asking other consultants to look at the (carefully anonymise d) behaviour sets provided by other consultants and invite them to comment on or evaluate these. These interview-like sessions were not recorded, but were carried out one-on-one with myself and notes taken. These exchanges (A few comments from this round are paraphrased below) provided tentative confirmation of a number of behaviours suggested earlier.:

NNS teacher (H)

H. can imagine that some people might [read the passage, filling in the easier deletions before reviewing the more difficult ones] in that way. H. feels this option is "astute" or "shrewd" (*schlau*)

NNS teacher (K)

K. agrees that it "makes sense" to apply knowledge of L1 syntax to 'grammatical' deletions. Cites closeness of many German and English structures

but apparent disconfirmation of others:

NNS teacher (K)

K. can't imagine that she would pre-read the (entire?) passage. Offers no reason why this should be so.

NNS teacher (R)

Thinks that filling blanks with L1 equivalents and then translating them can't be ruled out, but seems unlikely unless no other way of getting at the meaning is available.

One problem in this attempt was that some of these interviewees appeared to have difficulty in grasping what others had been trying to say. That this was not simply a question of linguistic difficulty is clear from the fact that the behaviour proposals on which they were asked to comment had been transcribed from audiotape and presented (edited for anonymity, clarity and economy but without, I think, doing violence to the meaning) in the consultants' L1. By my own informal assessment, a little over half of the behaviours offered by one individual could be mapped, without too much stretching, onto those of at least one other.

FUNCHAL

"The fairy lights of Funchal were spread out below us and the moon shone serenely down. The meat was roasted on open (1) FIRES/SPITS/GRILLS/BARBECUES before us, and while we ate, (2) A party of men and women from (3) THE hills sang and danced for (4) OUR entertainment. The words of the songs (5) THEY sang varied, but the melody was (6) ALWAYS/INVARIABLY the same; it had an elusive (7) QUALITY/RHYTHM/AIR , and it haunts us still. (8) WE from the little ships, making our (9) FLEETING/SHORT/RARE visits, see and hear things which (10) ARE not always vouchsafed to the ordinary (11) MORTAL/TOURIST/TRAVELLER, and we are grateful."

(SEMAC fillers provided byNS consultants shown in capitals)

'Intuited' behaviours

Shown below are the main behaviours or operations proposed in one form or another as useful in recovering FUNCHAL cloze deletions. The deletions to which consultants felt these operations could be relevant are numbered in the boxes. '>>' indicates textual links explicitly mentioned by consultants:

	Teacher A (NS)	Teacher B (NS)	Teacher C (NNS)	Teacher D (NNS)	Teacher E (NNS)
a	knowledge of TL collocations (1) "open fires" (7) "elusive quality (11) "ordinary people"	knowledge of TL collocations (1) "open fires" (7) "elusive quality (11) "ordinary mortal"	knowledge of TL phrases (1) "open fires"	knowledge of TL phrases (1) "open fires" (11) "ordinary men"	knowledge of TL phrases (1) "open fires"
b	knowledge of TL syntax (2) "a" (4) "our" (8) "we" (10) "are"	knowledge of TL syntax (2) "a" (4) "our" (8) "we" (10) "are"	knowledge of TL syntax (2) "a" (4) "our" (8) "we" (10) "are"	knowledge of TL syntax (2) "a" (4) "our" (8) "we" (10) "are"	knowledge of TL syntax (2) "a" (4) "our" (8) "we" (10) "are"
c		read beyond target sentence (9) "longer"		read beyond target sentence (9) "special"	
d		look back beyond target sentence (5) "a party of men and women" >> "they"	look back beyond target sentence (5) "a party of men and women" >> "they"	look back beyond target sentence	look back beyond target sentence (5) "our entertainment" >> "songs *we sang"
e	translate into L1 as needed (suggested 'vouchsafed'and 'elusive' as targets for translation)		translate into L1 as needed (11) (unsure of 'vouchsafed')	translate into L1 as needed (11) (unsure of 'vouchsafed', elusive')	translate into L1 as needed (11) (unsure of 'vouchsafed')
f	pre-read text (all?)			pre-read text (all or almost all)	
g		TL-L1 common item		TL-L1 common item ('barbecues' 1)	TL-L1 common item
h			L1 analogue/ common root	L1 analogue/ common root (suggested 'melody' 7)	
j	knowledge of root of TL item	m	knowledge of root of TL item (suggested 'elude' >>'elusive'		
k				"respond efficiently to problems"	
l					"Think of similar passages in the L1"
	intuition	intuition		"Knowledge" (example given was 'that travellers often get special treatment')	

Figure 4.1: cues & behaviours intuited as relevant to recovery in FUNCHAL

4.4 Selecting the final task passage

(1) Readability measures

The Flesch and Flesch-Kincaid indices (for a discussion of the former see Carrell 1987; Klare 1984) were used in assessing the 'mechanical' readability of potential expository passages (see below) for cloze task construction. I hoped by this means to quickly rule out passages which were markedly easier or more difficult than the norm, and these formulae were chosen in part because they were available as part of the software package GrammatikMac™, which at the time represented one of the few word-processor-compatible readability measures available for the Mac OS. A second factor, however, was Harrison's 1979 finding that Flesch-Kinkaid coincided better with expert judgements than other readability formulae. The GrammatikMac™ program factored in various surface features of a text, chiefly average number of syllables per 100 words, and average number of words per sentence, and provided objective 'reading ease' figure and grade level figures:

"MALARIA"
Flesch Reading Ease = 30.0
Flesch Grade = 15.9
Flesch-Kincaid = 17.2

The latter represents the approximate age grade within the US educational system at which the texts should be comprehensible to the majority of users. Given that my informants were non-native readers of English, I chose to aim a little below their chronological age of 18+, and looked for passages with a Flesch-Kincaid index of approximately 15–17. Alderson 2000 stresses that it is not easy to know just what features of a text create difficulty for readers, however, and for my

purposes here readability figures were never taken as more a rough indication of whether a passage might or might not merit further investigation

(2) Passage genre

Conventionally, passages used in testing comprehension are taken from the genre of expository prose (Heaton 1982), and (*pace* Oller's original implied claims for cloze) either selected or (Klein-Braley & Raatz 1985) revised to remove any over-specialised vocabulary, low-frequency lexis, etc. I conformed to these standards in making provisional selections of ten suitable passages, with the additional step of submitting these to three independent German L1 consultants (all education professionals) charged with (1) rejecting, with brief reasons, any passage(s) they felt German L1 undergraduates could not reasonably be expected to process and (2) marking any problematic elements in the passages they passed as otherwise suitable. I decided in advance that only passages accepted by two of the three consultants would be considered for use in data-elicitation, and that the same criterion would apply to any revisions I made to passage content. The FUNCHAL passage was, I felt, too short to be useful other than as a potential orientation/training passage (in which role it was later used) and so it was not regarded as a candidate elicitation passage.

Two rounds of this selection process provided me with six potential task passages out of the original ten, and the next stage was to mechanically delete every 7th word (the deletion ratio was set at 1:7 arbitrarily; see below) following the introductory sentences in order to create a standard cloze passage. This was

carried out initially via a home-made program on a Tandy 102 computer, and later by more reliable and flexible software available for the Macintosh OS.

4.5 Trialling cloze passages

Two criteria were identified to guide the selection of the eventual think aloud task passage from among the six candidates. Firstly, and I felt very importantly, could trial informants achieve a reasonable level of subjective success (rather than simply achieve a reasonable score): in other words, was the passage not so demanding that it would leave them frustrated or discouraged early in the task? Secondly, and of course critical, would the passage stimulate a worthwhile amount of think-aloud verbal reporting? I had already been warned by a researcher at another German university that, in his experience, it was extraordinarily difficult to predict whether or not a text would produce a sufficiently high level of introspection; this was the chief reason why I prepared six candidate cloze passages before proceeding.

Trialling passages: a conundrum

The six passages (average length 520 words, and comprising between 36 and 42 deletions) were each allocated to two GL1 informants (see table) working individually in a language lab setting.

	Informant 1	Informant 2	Informant 3	Informant 4
MALARIA	x			x
PLANTS			x	x
OLYMPICS	x	x		
FRIDGE		x		x
LUCID DREAMS		x	x	
EELS	x		x	

Figure 4.2: Informant/passage distribution in trials

As each informant was required to look at three passages, the task had to be spread over more than one LL session. Given that my teaching schedule sometimes conflicted with time-slots in which (a) informants could make themselves available and (b) times when LL facilities were under least pressure, each informant was provided with two sealed envelopes each containing one of her assigned passages and a blank C60 tape, thus allowing her to work on the task more or less independently. I did however make myself available as much as possible for consultation, and made a point of dropping in to observe informants at work whenever I had time. Informants were not given any special training in the task (for more on the training of cloze informants see chapter 5) but after a brief orientation in the nature of think-aloud, they were simply instructed to

(1) do the cloze task as though it were a test while saying as much as possible about how they were finding filler words, any comprehension difficulties they experienced, etc. They were also asked to

(2) give, at or near the end of each passage task, an assessment of how difficult they had found it overall. An informal scale (glossed in the

informants' L1 for clarity) was proposed of 'easy'. 'a little difficult', 'pretty difficult' and 'extremely difficult', but informants were not obliged to stick to this. The only other instruction that I can recall giving to all informants was that

(3) they should try not to leave any passage deletions unfilled.

In these circumstances it was not possible to maximise the quantity of verbal report by way of reminders to talk.

The passage entitled LUCID DREAMS (Flesch-Kincaid 17.8) was described as extremely difficult by both trial informants, whose scores fell below SEMAC 50%. The passage ELECTRIC EELS (Flesch-Kincaid 15.2), on the other hand, was felt to be rather easy by both of the informants who worked on it, and whose scores were over 80%. This left four passages, all between 16.6 and 17.4 on the Flesch-Kincaid scale, and on all of which informants had been able to score between 58% and 74%—a range thought to be suited (Bachman 1990) to a test procedure, and thus appropriate in a data-elicitation task intended to resemble a test in some respects. These were MALARIA, focusing on the nature of the disease and the discovery of a treatment for it; OLYMPICS, about the ancient games; FRIDGE, on the topic of the dangers posed by inappropriate food storage; and PLANTS, which discussed how plants allegedly communicate with one another. All of the passages could be described as fairly regular examples of expository prose such as make up a good part of the reading diet of ESL (or at least EAP) students.

From the trial sessions which I had been able to observe it appeared that, despite their surface similarities, some passages appeared to stimulate more verbal

reporting than others. Time pressure forced me to gauge only very roughly the amount of think-aloud reporting which each informant had produced by taking ten extracts of thirty-seconds (or, rather, thirty units on the tape counter) of each of the eight informants' protocols, and impressionistically evaluating the amount of verbalization each extract contained as 'very high', 'high', 'moderate' 'low', or 'very low'. The first extract was audited beginning at ca. 150 on the tape counter (i.e. approximately two and a half minutes into the protocol), and at the end of the thirty unit/second extract the tape was wound on for thirty units and a second extract was audited.

This process was continued (and in order to avoid biasing my judgement I took care not to audit in sequence the protocols of informants working on the same passage) until ten extracts had been evaluated, i.e. sampling a span of six to seven minutes of each protocol. Where an extract contained effectively *no* interpretable verbalization it was abandoned and the tape wound on for thirty units and a new extract audited. This happened on only three occasions.

Each extract audited was scored on a 1—5 scale according to how much reporting it had contained, with '5' corresponding to 'very high' and '1' to 'very low'. The total score for each passage was calculated by adding the two informants' scores for each extract and taking the average, rounded here to the nearest decimal point, e.g. Informant P./MALARIA

Extracts

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	Av.
2	3	2	1	2	3	2	3	3	2	*2.3*

The resulting figures were:

OLYMPICS = 3.4

MALARIA = 2.8

FRIDGE = 2.6

PLANTS = 2.5

Figure 4.3: Ratings of passage productivity in trials

The clear lead of the OLYMPICS passage over the other candidates remains unexplained: its difficulty as measured by the readability index (This refers to the full version of the source passage, which in use had to be curtailed for practical reasons.) or the test-takers' SEMAC scores was not appreciably different. Nor, when I asked them to quickly look over each of the three passages they had not yet seen, did informants identify OLYMPICS as conspicuously more, or less, difficult or interesting.

The opportunity arose a week or so later to confirm this result. I set each student in a group of twelve high-intermediate learners of English (in an adult evening course) either the OLYMPICS or MALARIA cloze task, under roughly similar conditions to the undergraduate trial informant group, and also in a LL classroom. These learners were invited to make a recording of themselves carrying out the cloze task, and to later audit this recording to find out which passage clues or other information they had used in filling the blanks. The up-to-date LL facilities allowed me to divide the class into two groups for monitoring purposes, and to quickly evaluate the amount of verbalization going on by switching among

individual participants and checking an LED display showing the 'signal' level each was generating. As even 'noise' in the form of superficially meaningless tongue clicks, humming, etc. would also have generated a signal on the display, I obtained the group's permission to listen in briefly to participants' actual recordings as well.

As each participant in the session (which lasted, depending on the individual between 20 and 30 minutes) was 'eyeball' monitored via the signal level display at least ten times, with each visual check lasting three seconds, and separately audited at least five times, with each sampling lasting approximately five seconds. Each participant was thus 'sampled' at least 15 times in all, in an evaluation procedure which can perhaps best be described as systematically impressionistic.

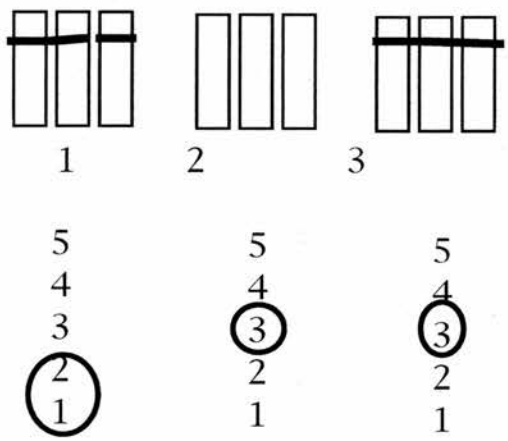

Figure 4.4: Format used in recording informant verbal output

The diagram above shows part of the sheet on which I recorded my impressions of individual participants' verbal output. The boxes above—only the first three of ten are shown—echo the format of the LED display, while the number columns

below represent a scale of verbal production. A number scale was chosen for the audited samples in the expectation that an average figure might later be calculated (although this later proved superfluous) while a rough impression seemed adequate for the display samples. The diagram shows that this participant's visually sampled output was high in the first and third checks, but absent in the second. His or her audited output was low in the first sample, moderate in the second, and moderate in the third.

Review of the note sheets completed in this session confirmed that here too the OLYMPICS passage seemed to stimulate more (and, based on the visual sampling, more sustained) thinking or talking aloud than the MALARIA alternative. I decided at this stage that I had enough evidence of the OLYMPICS passage's superior productivity (even if I still had no real clue as to the underlying reasons for this) and resolved to use it as the primary stimulus in future data-elicitation sessions. Although some words (i.e. potential blanks) in the passage arguably a degree of specialise d knowledge of sports events (discus, pentathlon, etc.) or were fairly low-frequency items (waived, laurel) I decided against revising the source passage. I was, after all, interested in how informants dealt with deletions across the full spectrum of difficulty.

4.6 Selection of the stimulus text for JL1 informants

The possibility of requiring data from Japanese L1 informants had not yet arisen at the stage of planning for data-gathering among German informants. When the time came to gather such data from Japanese, I gave careful thought to whether to

employ the same OLYMPICS text as the task passage. Some of the arguments for and against such a step are worth setting out here.

Against the use of the same text, it can be argued that, first of all, the overall level of English proficiency of the Japanese sample was slightly lower than that of the German sample. It should be noted in this context that the available comparison measures of (in Germany) the Berlin university's and (in Japan) university A's and B's placement examinations were far from being equivalent measures. Moreover, the TOEFL and TOEIC tests, taken by many Japanese students, are relatively little used in Germany, so that no 'objective' measure (see below) was available. Secondly, it was intuitively likely that knowledge of the ancient Olympic Games—as opposed to the modern games—and indeed of other aspects of ancient Greek civilisation, were more likely to feature in the background knowledge of German (undergraduate) informants than Japanese. (This, as it turned out, was not entirely the case.) Finally, English and German are almost incomparably closer in terms of structure than are English and Japanese. There might then be a case for setting Japanese informants a stimulus task rather lower in linguistic challenge.

In answer to the first point above, it must be noted that a number of my potential Japanese informants were 'returnee' students, who had spent some part of their lives in the UK or the USA (or in any of a variety of other countries) As many of these informants had been educated in either the UK or US school systems, or in international schools in which the language of instruction was English, their L2

oral/aural skills were often at a high level. (The same could generally be said for their reading skill, but by no mean always for writing.) One or two of these, indeed were close to native-speaker levels of ability. There was every chance that such individuals would form part of the JL1 informant pool.

I was subsequently informed, by Japanese secondary school teachers and Japan-educated acquaintances, that anyone educated in the Japanese school system can be presumed, with near certainty, to have at least been *exposed to* information about the ancient and modern Olympic Games as part of her school curriculum. This stems in part from the popularity of sports in Japan, but is mainly due to the central historical role of the 1964 Tokyo Olympics in Japan's post-war rehabilitation. But perhaps the most persuasive arguments for retaining the OLYMPICS passage as the primary tool was that it had been shown to work reasonably well as a stimulus for reporting so far. Due to an unusual coincidence of circumstances, moreover, I found myself with a teaching load of just over twice the regular number of classes and so had very little time to devote to the search for a replacement task passage. Finally, if my intent was to compare the cloze processing behaviours of German L1 and Japanese L1 informants, a common task was highly desirable.

4.7 Choosing a deletion procedure

I chose to work with the 'standard' fixed-ratio, (pseudo)random cloze deletion procedure. This may seem perverse, given that I was focusing on test-taking processes and in particular on how cues are activated: introspection about

'content' items is (I found in a 1993 study) typically more informative than that for 'function-word' deletions. Why not, then, use a rational deletion procedure, which could be so designed as to target content words consistently?

The reasons for selecting a fixed-ratio deletion procedure were fourfold. First of all, fixed-ratio deletion appeared still to be the most common cloze deletion method world-wide and hence the procedure closest to what students might already have encountered. (And this situation persists within ESL in Japan: over two-thirds of the word-level gap-fill passages I know to have been used within the last 24 months by teaching colleagues, as didactic tasks and/or assessment measures, have used fixed-ratio deletion.) Secondly, the range of deletion procedures which have been labelled as 'cloze' (see Chapter 2) is extraordinarily wide, embracing most forms of textual mutilation imaginable.
Cross-comparability among studies using different deletion methods may be limited (See Oller & Jonz 1994:371ff) and so it seemed useful to return to an easily-defined and easily-replicable deletion procedure. Thirdly, and despite the poorly-supported claims to the contrary reported in chapter 2 a (pseudo)random deletion procedure like every *n*th word appears to target words objectively, and thus requires the fewest assumptions about which sorts of item test-takers might experience difficulty with, or—a related point—to require the least justification in terms of where that elusive quality 'redundancy' is believed to lie. Finally, I was also interested in seeing what information could be gleaned about the processes involved in 'automatic' recovery of deletions. Function-word cloze items often

appeared, in self-as-subject trials and the protocols of early German think aloud informants, to be recovered very quickly indeed and with very little or even no attendant verbal reporting. The German article system (to take just one functional-item class) differs from that of English in some ways, but there are many parallels. Japanese, however, lacks an overt article system, and it would be interesting to see how the recovery of such items by Japanese L1 informants would compare with that of their German counterparts. A deletion procedure which targeted function as well as content items was thus called for.

Another factor that coloured my thinking about the choice of deletion mode was the widespread view (cf. Read 2000; Alderson 2000) that cloze is better employed as a test of overall language ability than of reading ability. (For the contrary position see Jonz 1976, Aitken 1977.) Some attention to structure or function items was thus indispensable. I will have little more to say on the subject of what I believe cloze to measure, or what its key construct(s) might be. I cannot claim to know what cloze measures. Moreover, one of the implications of the alleged unpredictability of natural cloze (Alderson 1978; Porter 1978; Klein-Braley 1981) may be that that in terms of real-world practice it makes more sense to conceive of it as a set of individual measures than as a unified whole. Firm definition of the construct(s) of cloze I leave to others and the future. My goal in gathering think-aloud and other data from cloze task-takers was to sample the kinds of thing people do when faced with this (for better or worse) common task format, and how often they do them.

4.8 Scoring cloze success

The two main criteria in scoring cloze test-takers' performance, 'exact-word' recovery and 'semantically acceptable (SEMAC)' recovery, were touched on in chapter 2. These need, however, to be elaborated here in background to my decision as to which scoring method to adopt. Oller 1976a recommends exact word scoring "for global proficiency testing purposes" on the grounds that this is easier to use. An obvious counter-argument to the exact word criterion, however, is that it measures the test-taker not just against the standard of a native speaker of the TL, but against one particular TL native (i.e. the author of the passage) and that at a given point in that individual's life. If we aim to measure real-world TL comprehension and production through cloze it makes better sense to accept as correct any filler which a typical native speakers would find acceptable. There is, anyway, evidence (McNamara 2000:16) that the *ranking* of test-takers' scores does not significantly change whether the exact or SEMAC criterion is used, and in favour of SEMAC scoring some evidence (Lange & Clausing 1981) that it produces slightly better discrimination among test-takers. There appears to be a rough consensus among testers (cf. Bachman 1990) that semantically-acceptable (SEMAC) rather than exact-word scores are to be preferred when scoring cloze tests carried out by non-native speakers of the target language, although Alderson 2000:207 merely notes that the two criteria exist. It may also be that a cloze passage scored by the 'exact word' criterion requires a larger number of items: Sciarone & Schoorl 1989:415 claim that 75 items suffice with SEMAC scoring, but 100 are required with exact-word scoring. The question of minimum item

number in cloze appears (cf. Alderson 2000) to be open, however.

My experience of using cloze in the classroom as a didactic task, perhaps as important as any of the points above in forming my opinion, has consistently been that second-language takers of cloze see the exact word criterion as "unfair", and find it difficult to see why they should *not* receive credit for supplying what the test-giver may have to admit is a semantically and contextually acceptable replacement for the missing word. Given that I had agreed to make individual informants' scores available to them, along with explanation of why answer *x* was right or wrong, there might be scope for friction with informants were the exact-word criterion to be used. As it was likely that the willingness of future informants to come forward would depend to some extent on how acquaintances and classmates had felt about the experience of reporting, a affective response was to be avoided.

German and TL L1 consultants' SEMAC fillers

SEMAC scoring it was to be, then, and my first step was to elicit from non-native speakers as well as native speakers of English a set of semantically- appropriate filler items. I gathered these from both NNS and NS speakers for two reasons. Firstly, it seemed important to try to form some picture of the kinds of answers which would be considered acceptable by the group (i.e. non-native speaker teachers) at least statistically most likely to set the cloze tests students might actually encounter in Germany and Japan, as well as those considered acceptable by native-speaker 'experts'. Secondly, I wished to see whether one set of answers,

that of the NNS or the NS group, conformed more closely to the answers volunteered by NNS informants. My working hypothesis was that adult non-native speakers from the same language background as the informants would be the better predictors of the responses these would make.

To this end five German-L1 consultants (all either trainee, working, or 'resting' educators) were individually invited to contribute fillers for the OLYMPICS cloze passage. All were familiar with cloze, and with the notion of semantic acceptability, which was operationalise d here in terms of 'making sense in the context of the passage' without taking (atextual) historical accuracy as a major criterion. By this measure, laurel, bayleaves, etc. would all be possible fillers for deletion (28), for example. Consultants were instructed to give first the filler they thought most appropriate. They were also invited to suggest any alternative fillers which they thought could genuinely fill the blank, but to mark these on a separate line on the sheet provided and, regardless of the order they were noted down in, to number them in descending order of perceived semantic appropriateness. The items in boldface below represent the kinds of 'write in' information required:

some official (7) records date from 776 B.C.

documents (3) **histories** (2)

In response to consultants' questions, I added the following instructions: (1) if they could not suggest a TL filler then they should note down any L1 equivalent which seemed to fit the blank semantically and (2) if they noted down a TL filler of whose meaning they were unsure, they could add an L1 gloss alongside it. I did

not anticipate that an informant would in fact enter a filler which she did not feel reasonably confident about, and in the event none did. Experience to date had suggested, however, that it was as well to take on board informants' 'what if' questions and provide some means of satisfying these.

German L1 consultants produced a considerably wider range of fillers for 'content word' deletions such as (20), (22), etc. than I had anticipated, and by no means all seemed appropriate to this NS evaluator. This is unsurprising, for although non-natives might be expected to possess smaller active vocabularies in English, and hence produce fewer candidate fillers, they are also less likely to be aware of the fine distinctions of semantics, collocation and context that drive lexical choice among native-speakers. They are thus quite likely t0 include candidate fillers which a native-speaker would find unacceptable.

The set of fillers thought acceptable by NNS consultants alone was gathered for comparative purposes only, and was not intended to serve as as a guide for SEMAC scoring. For this I wanted the considered opinions of English native speakers, and this led me to my next step.

TL native-speaker consultants' SEMAC fillers

Five native-speakers of a range of regional varieties of English, American, and Australian—were invited to review and edit the compiled list of candidate fillers produced by the non-native group, but chiefly of course to suggest candidate fillers of their own. These consultants (of whom two were teachers of EFL, one a researcher in psychology, one a translator, and the last an airport baggage handler)

worked independently, and when I collated the final list of SEMAC fillers I excluded any which had not been suggested and/or approved by at least three of the native-speakers. A number of (to my mind) inappropriate fillers—not all of which had originated with non-native consultants—were removed via this procedure. I have made no changes to the list thus compiled, except to remove one item on the (I think objective) grounds that the passage punctuation—specifically, the lack of one or more commas—renders it objectively inappropriate. This was 'moreover' in deletion (17). Despite my personal unease over a few items in the list below, such as 'ball' and 'hammer' in deletion (21), I made no further changes. The stages outlined above were to some extent forced on me by my own others' schedules but it is clear with hindsight that by allowing NS consultants to see the fillers produced by their NNS counterparts (and thus, as some did, to base their own lists of fillers on these) I unwittingly confounded the two sets, thus making it unrealistic to compare and contrast the choices each group might separately have made.

Consensus SEMAC fillers

Native-speaker 'consensus' SEMAC fillers for the OLYMPICS cloze passage are shown below. The original deleted item is shown in italics in column two, with the SEMAC alternatives following. No other fillers conforming to the single-word rule were proposed by more than two of the five NNS consultants, and so the items shown below also represent a consensus on the part of both NS and NNS respondents. As mentioned above, it seems superfluous to isolate here those fillers

proposed only by NNS consultants.

Del	Passage context	Exact word + SEMAC fillers approved or proposed by NS consultants
1	In ancient Greece athletic festivals were very important and had strong religious associations. The Olympian athletic festival, held every _____	*four,* (any integer), few
2	years in honour of Zeus, eventually _____	*lost*, relinquished
3	its local character, became first a _____	*national*, Greek
4	event, and then, after the rules _____	*against*, banning, forbidding,
5	foreign competitors had been waived, international. _____	*no*
6	one knows exactly how far back _____	*the*
7	Olympic Games go, but some official _____	*records*, documents, inscriptions, histories
8	date from 776 B.C. The Games _____	*took*
9	place in August on the plain _____	*by*, of, near, below, before, beneath, beside, around, surrounding
10	Mount Olympus. Many thousands of spectators _____	*gathered*, came, arrived, flocked, descended
11	from all parts of Greece, but _____	*no*
12	married woman was admitted even as _____	*a*
13	spectator. Slaves, women and dishonoured persons _____	*were*
14	not allowed to compete. The exact _____	*sequence*, order, schedule, number, range
15	of events is uncertain, but events _____	*included*, comprised
16	boys' gymnastics, horse-racing, field events _____	*such*
17	as discus and javelin throwing, and _____	*the* some
18	very important foot races. There was _____	*also* always
19	boxing and wrestling and special tests _____	*of*
20	varied ability such as the pentathlon, _____	*the*
21	winner of which excelled in running, _____ ,	*jumping*, swimming, ball, hammer
22	discus and javelin throwing and wrestling. __	*the*

23	evening of the third day was _____	*devoted*, dedicated, allotted, assigned
24	to sacrificial offerings to the heroes _____	*o*
25	the day, and the fourth day, _____	*that*, day, night
26	of the full moon, was set _____	*aside*, apart
27	as a holy day. On the _____	*sixth*, fifth, next
28	and last day, all the victors _____	*were*
29	crowned with holy garlands of wild _____	*olive* [original passage item], laurel, bayleaf, flowers
30	from a sacred wood. So great _____	*was*
31	the honour that the winner of _____	*the*; a
32	foot race gave his name to _____ year of his victory!	*the*

Figure 4.5: Consultants' elicited SEMAC fillers

Exception may be taken to some of the consultants' proposals, and my own feeling at the time was that these seemed unusually generous in regard to certain deletions. The list did, however, appear preferable to one based solely on my own intuitions. It may be worth mentioning at this point that similar procedure to that outlined here was attempted with Japanese L1 consultants, in an attempt to see whether and how their sets of candidate fillers might differ. This was brought to a premature end, however, when I noted (despite a clear request to work independently) the consultants had begun to collectivise their efforts. I thus used the same set of SEMAC fillers, above, with Japanese informants.

Scoring by item class

Success on 'grammatical word' deletions (idiosyncratically defined—see below) was calculated separately from that of 'lexical word' (or 'content word') deletions, on the grounds that the former may be seen an index of the testee's grasp of

English structure and the latter a measure of her lexical knowledge. My working hypothesis was that informants who scored poorly on grammatical word deletions would be unlikely to do well in recovering content words. Those who did well on lexical deletions. however, would be more likely to do well also on grammatical word blanks.

In some cases the phrasal setting of the deleted item had to be taken into account in allocating it to one category or the other, as in (25) 'set aside'. This was classed as a lexical item on the basis that it is part of a phrasal verb, in which form it would have been acquired by informants. Note also that I have deviated from the criteria for distinguishing grammatical and lexical words set out in Finch 2000:135 by including prepositions, e.g. in (4) and (26), in the class of lexical deletions. This decision was informed largely by the existence of plausible alternative fillers for each of these items. The table below shows items listed as grammatical/ functional and lexical/content within the 32 deletion OLYMPICS passage.

'Grammatical' deletions	'Lexical' deletions (exact word only shown here)	Grammatical' deletions (cpnt.)
(5) No	(1) four	(25) that
(6) the	(2) lost	(28) were
(9) of	(3) national	(29) olive
(11) no	(4) against	(30) was
(12) a	(7) records	(31) the
(13) were	(8) took	(32) the
(16) such	(10) gathered	
(17) the	(14) sequence	
(18) also	(15) included	
(19) of	(21) jumping	
(20) the	(23) devoted	
(22) the	(26) aside	
(24) of	(27) sixth	

Figure 4.6 Structural vs. lexical deletions in OLYMPICS

4.9 Recruiting German L1 informants

The initial rounds of data-gathering employed German L1 undergraduate informants from a Berlin university at which I was teaching. Students from intact EFL groups were invited to volunteer, without payment, and some of these individuals brought along fellow-students from other English courses who had expressed a willingness to participate in the data-collection. Recruitment of informants was thus partly by direct solicitation and partly via 'snowball' recruitment. Informants from intact groups which I myself had taught were considered the most desirable, as I was inevitably better acquainted with their personalities and English ability-levels. I could, moreover, call on teaching

colleagues who also taught these individuals for their evaluations. Due to the limited pool of intact group volunteers, however (and so as not to antagonise those who wished to self-select a partner) other volunteers were welcomed.

A few idiosyncrasies of individual informants may be worth mentioning. I noted that those who expressed the preference for pair-condition reporting (see chapter 6) volunteers from the former DDR often paired off with and even specially recruited, other former easterners. This was taken to reflect the, at that time, still quite perceptible gaps in attitudes and expectations between students from western and eastern backgrounds. Two informants from southern Germany also brought along partners from their own part of the country, and this was also welcomed on the grounds that the more comfortable informants felt in the reporting situation, the more productive they were likely to be. It was, for example, anticipated that informants might wish to use their own regional dialects in reporting—a phenomenon which did appear to a minor degree. It was noted that women were more likely than men to actively express a preference for pair condition reporting, but I saw no evidence of a marked preference for a partner of the same (or indeed the other) sex. Given the preponderance (ca. 70% on average) of women in the courses from which most informants were drawn, however, it was inevitable that most informant-pairs would consist of two women. As 75% (n=18) of the volunteer informants were women, the gender imbalance was even more pronounced.

The mean age of German L1 informants was approximately 21.5 years. Due to the need to wait up to several years before being allowed to enrol in particular courses at certain universities (the so-called *numerus clausus* system) German undergraduates as a whole are slightly older than their peers in other countries. The until recently fairly relaxed regulations governing the permitted period of study also allow undergraduates to remain at university for much longer. The youngest GL1 informant in my sample was 18, the oldest 34. The amount of time (in months) which informants had spent in an English-speaking environment (ESE) was recorded. The German think aloud L1 informant pool (whose output has been included in the data reported here) can be broken down as follows:

German L1 think aloud informants n = 24
Male 25% Female 75%
Mean age = 21, range = 18–34
Time spent in ESE: mean = 2 months, range = 0–36 months
Number of solo condition informants = 8
Number of pair condition informants = 16 (i.e. 8 pairs)

4.10 Japanese L1 informants

It may be as well here to describe the recruitment of the later groups of Japanese L1 think aloud informants. Japanese undergraduates have typically proceeded directly from school to university, although a number may have studied for one or more years in *yobiko*, special cram schools which tutor students in the often markedly idiosyncratic entrance examination questions favoured by the particular colleges they hope to enter. Most undergraduates at the two Japanese universities at which I was able to gather data had passed their respective college faculties'

entrance examinations. These exams are considered to be among the most demanding in Japan. Some informants, however, had risen through the so-called 'escalator' system, by which students of primary (elementary) or secondary (junior and senior high) schools affiliated to the university in question are allowed to enrol as undergraduates via an alternative entry procedure. There tends to be a certain, if you like, *esprit de corps* among undergraduates who have passed the entrance exams, and a concomitant reluctance among some 'escalator' entrants to publicise their status. I do not distinguish in reporting data between 'escalated' and 'examined' informants here, chiefly because the distinction has little real bearing on the focus of the study. Neither sub-group of informants was (on their own reports) significantly more likely than the other to have had prior experience of cloze outside my courses or those of other teachers, or (on the basis of classroom observation and test results) to be much more proficient, overall, in English. 'Examined' students had learned the skills, language and metalanguage needed to answer the traditional translation-heavy, structure-centred questions in the entrance exam, but might find it difficult to function in more authentic language contexts. 'Escalated' informants had (being more likely to have spent time overseas) typically had more experience of 'communicative' or 'naturalistic' English learning. Language teaching in Japanese schools is still so heavily oriented towards the older paedagogical methods, however, that the fruits of such experience may be markedly degraded by the end of secondary education.

Given that both universities A and B are prestigious institutions, their affiliated schools tend to attract a higher proportion of children of successful and/or wealthy parents. I was aware that undergraduates from such a background were inherently more likely to have spent time overseas—which might have some considerable impact on their task performance—and so (as I had done by GL1 informants) I recorded the number of months informants had spent in an English-speaking country or school. This averaged 4 months, with a range of 0 to (an extreme outlier) 60 months.

Japanese think aloud informants *n* = 24
Male 17% Female 83%
Mean age = 20, range = 18–21
Time spent in ESE: mean = 4 months, range = 0–60 months
Number of solo condition informants = 8; Number of paired informants = 16 (i.e. 8 pairs)

The figures above include think aloud informants only and do not include the 'postal survey' GL1 and JL1 respondents discussed in chapter 6. Moreover, additional JL1 informants were recruited for investigations of alternative data-elicitation procedures (so-called 'non-continuous reporting' and 'annotated cloze'.) These smaller informant samples are described in chapters 6 and 7. Some small-scale surveys related to this area of study were conducted at various times, using either 'didactic' tasks with intact class groups, or focus group-type discussions with invited individuals. These too are described at the appropriate points.

4.11 Comparability of informants

Reliable comparison measures of think aloud informants' English proficiency were not always available, but based on the following data I felt able to build up a reasonable picture of the range of abilities represented in the informant groups.

German L1 informants

All of the Berlin German L1 informants whose data is included in this study had passed an interview procedure in order to gain admission to their university's *Eingangsphase* English/*Leistungskurs* (Entry stage English/Intensive Course) and of course all had acquired the German *Abitur* or secondary school-leaving certificate. All of the informants from this group had participated in at least one semester of English classes taught by myself and/or one or more colleagues (whose assessments of the informants ability levels were also well-grounded on classroom observation and test scores) so that the 'thick' picture I had of informants included their observed classroom behaviour, their (apparent) level of confidence in their English ability, and of course of their task-performance in the primarily communication-oriented lessons. In addition, it had been possible to make informal assessments of the informants' 'passive' grammatical and lexical knowledge through classroom exercises, tests and activities. Although no formal/objective assessments of these aspects of performance were made, it is possible to say that the informants whose data is included in this study appeared to be of roughly comparable competence in terms of knowledge of English grammar and lexis. Differences certainly existed in the facility and accuracy with which individual informants were able to produce or actively use English, but I noted

(and this was confirmed by a colleague who also taught the same students) that individuals' productive performance appeared to vary considerably according to the task in hand. Some students performed better in role-plays or oral interactions, while others wrote with greater facility or efficiently extracted and orally presented relevant information from texts. My own impression was that students from southern and eastern Germany were slightly less confident in oral interaction, but compensated for this with a higher level of grammatical knowledge and reading and writing skills.

GL1 informants also took one of a small pool of sample 20 deletion cloze tests during a trial think aloud session, and scores on these tests were recorded by myself. The few potential informants who asked about their scores on the test were told it was "just above average", whether or not this was in fact the case. My justification for this economy with the truth was threefold: I did not want to lose a single informant through the demotivation that a lower score might induce; my interest lay not in each individual's test product, as it were, but in the processes by which she completed the test; the range of individual scores on these orientation tasks was anyway not very wide at all (overall mean 70%, range 65%—80%)

Japanese L1 informants

All of the Japanese L1 informants whose data is included in this study had passed either the entrance test for the university A's intensive English course (as well as the demanding entrance examinations to one of that institution's academically highest-ranked faculties) or the examination for admittance to university B's

highest-ranked faculty. Given the standing of these faculties, only the most able 'escalated' students of affiliated high schools were allowed to enroll there. Colleagues at both institutions and contacts elsewhere have confirmed that there should be only fairly minor differences in ability between informants from the two groups. I do not, therefore, identify informants by their college affiliation.

All of the informants from this group had participated in at least one semester of English classes taught by myself and at least one colleague, so that a second opinion was available in rating each informant's English proficiency. As with the GL1 group, I was able to build up a fairly detailed picture of each informant's classroom behaviours, her (apparent) confidence level, and of her task-performance in the essentially communication-oriented lessons.

All JL1 informants took one of the same pool of short sample cloze tests during a trial think aloud session. Scores on these cloze tests were lower overall (overall mean 65%, range 50%-75%) for the Japanese group than for the German group and the range wider. Essentially the same answer described above was given, and for the same reasons, to JL1 potential informants who asked about their test scores. Where these were available, i.e. for approximately half of the individual informants, I was able to roughly match scores on TOEFL, TOEIC, or the Japanese STEP (*Eiken*) tests with my own impressions. Precise correlations were not attempted, but my own rankings of informants on a five point impressionistic scale fitted acceptably with the available test scores and the impressions of other teachers where these were available. I already knew that university entrance

examination scores would be unavailable for purposes of comparison, and so made no attempt to find these out.

Other comparators of ability

No comparison was possible between the university entrance exam scores of my German sample and those of potential Japanese informants, but as students at Japanese universities almost all know their current TOEFL status very clearly, it appeared that TOEFL might provide a rough-but-useful comparison measure. From a colleague in the university faculty from which my German L1 sample was drawn I obtained an informal cross-referencing of scores on that university's internal English tests with institutional TOEFL scores. This suggested that a TOEFL score of around 460 represented a minimum level at which Japanese L1 informants could be thought to have a level of English ability roughly comparable to that which German L1 think aloud informants had demonstrated in their university's tests. All Japanese informants claimed to have met this criterion either on an actual TOEFL exam, or on one of the regular practice sessions held in Japanese cities.

Other and perhaps more appropriate comparison measures were also taken into account in comparing the TL proficiency of Japanese and German potential informants. These were (1) the respective group's performances on certain classroom tasks which had been set to both German and Japanese groups, such as reading comprehension tasks, 'didactic' cloze and other lexical inference tasks, and gap-filling exercises, and (2) my own instincts as an experienced classroom

teacher. The manner in which German and Japanese students approached the abovementioned classroom tasks diverged in some respects (use of the TL; time required, etc.) but their overall levels of task success were felt to be comparable for my purposes here.

4.12 Informant orientations to think-aloud (GL1 and JL1)

The task required of informants was explained to potential volunteers in English and in their L1, and this verbal explanation was accompanied by a written explanation in English as well as German or Japanese, as appropriate (See appendix 1 for a TL version of the explanatory handout.) Questions were then invited in English or the informants' L1, and these were answered with reference to the content of the written explanations. Questions common to both groups included: "What kind of test is this?", "What are you trying to find out?"', and "Will we be told our scores?" These were answered respectively with "You tell me."; "What people are doing when they work on this kind of task, and how they find the answers.", and "Yes, but only if you come and ask me in person. I won't give out scores to the group."

Another point which potential informants from both language backgrounds wished to have clarified was whether I had a preference for verbal reporting in English or in the L1. Anxious to bias informants' choice of language or reporting (LoR) as little as possible (and, later, because I was aware that some potential JL1 informants had noted that my command of their native tongue was very far from perfect) I assured them that I had access to help from native speakers of German

(and Japanese) so that they should not avoid using their L1 for my sake. The key point that I adhered to in my responses was that informants should report in whichever language they were thinking in *at the time*. The language of reporting (LoR) might, I emphasised, sometimes be English and at other times the L1. There was no preference whatever on my part for one LoR over the other.

Following my verbal orientation, I set participants a think aloud cloze task which required then to verbally report into a microphone as they worked on a short cloze passage (FUNCHAL, THE BLANKET, etc.) Both the initial German and Japanese orientations were held in a language laboratory (or, for those who could not attend the main sessions, in an office or classroom) so as to provide a sufficient degree of privacy to allow informants to think-aloud without embarrassment. Observation and auditing of solo-informants strongly suggests benefits, in terms of verbal output and affective response to the task, to having them use headphones so that they can hear only themselves, and if possible to screening them from the eyes of any other participants. This 'privacy' effect has been noted in both GL1 and JL1 groups, and is supported both by the retrospective comments of orientees and by the results of a questionnaire given to JL1 classroom students who had listened to a short lecture (in English, and intended as a listening comprehension exercise) about current applications of think-aloud in usability and other research. Selections made on the questionnaire revealed that over 80% of respondents though it 'important' or 'very important' that "Others should not be able to hear what I am saying as I think out loud."

4.13 Selection among candidate informants & 'low verbalisers'

Orientation sessions were intended to clarify the basic idea of think-aloud and to give participants the hands-on experience which would allow them to decide whether to continue as informants. I often found myself under some pressure to accept more or less on the spot those who wished to enroll as informants, and this inevitably made it difficult to properly audit recordings with a view to filtering out or dissuading weak reporters. Some orientation participants did seem to realise that thinking-aloud was not their strong point, and withdrew themselves from further participation, while others acknowledged that they had not produced much data in this initial session but felt that they would do better in future. This was entirely plausible (cf. Ericsson & Simon 1993:xliii), and for this reason as well as that below I accepted a number of informants for whom the think-aloud task turned out to be inappropriate (or vice versa.)

Whether this was due to the unfamiliarity of the task, or the amount of time it was expected to require, I experienced considerable difficulty in recruiting informants and scheduling data-gathering sessions around their often changing part-time work commitments; the perpetual shortage of informants made it essential to use as many as possible of those who had volunteered, even (as mentioned above) where these did not seem at first to be entirely suited to the task. This may be a suitable point at which to discuss the variety of informant I have called the 'low verbaliser.' These informants were, not unsurprisingly, encountered only among solo informant, but it is less obvious (though see below) why they should have been more common among male informants and in particular among Japanese

male informants. The four JL1 low-verbalisers, in fact, were all male, although the task peformance of some female informants (cf. 'Harumi' in chapter 5) approached the criterion, below. Three low verbalisers—two male–were identified among German L1 volunteers, but of these two withdrew following orientation and practice. The remaining male informant felt that he would perform better on the next task, and this was accepted.

The criterion for labeling individuals as 'low verbaliser' was conservative: they had verbalized so little that it was possible to interpret their processing behaviour on no more than four items (i.e. approximately 10%) of the deletions in the passage. No marked differences set off low verbalisers from their more productive peers in relation to comparison language tasks such as didactic cloze, lexical inferencing, reading comprehension, etc. Some, it is true, appeared a little hesitant in pair- or small-group work in class, but this was did not prevent them from completing the set tasks. A simple lack of gregariousness cannot explain these individuals' poor performance in think-aloud, for perhaps only self-directed, silent thinking could require less human interaction (but see below.) Poor language skills in the L1 do not explain it either, for in almost all cases the little reporting low-verbalisers achieved was carried out in the L1. Nor, given that their performance in other tasks was not inferior to that of their peers, does lack of 'background knowledge' appear a likely factor, and when two JL1 low-verbalisers were given the opportunity to think-aloud about other passages (choosing from a range of passage topics) their performance typically remained weak. This is in

contrast to the majority of thinkers-aloud, whose volume of reporting (though not necessarily its 'quality') tended to improve from practice task to data-elicitation task.

It is hard to tell how often low-verbalisers are encountered in studies employing think-aloud or other forms of verbal report for, reasonably enough, few writers raise the issue of informant-productivity. The sweeping claim in Ericsson & Simon 1993:224 that:" "Thinking aloud and talking aloud can be elicited almost instantaneously by the appropriate instruction from virtually all human adults" does little to encourage attention to the point, moreover. Afflerbach 2000: 168 seems to acknowledge the issue, however, in his comment that

"The majority of protocol analysis research focuses on talented readers [...] who are often more verbal [...] and may better verbalise the things they do in a think-aloud."

The same author (op.cit.:179) suggests that verbal reports appear to be closely related to the notion (Vygotsky 1978) of 'inner speech' which can take the form of an intimate and potentially revealing conversation with oneself. Part of the challenge of think-aloud may lie in editing the internal conversation so as to render it externalisable, and this may be to some extent a psychological-linguistic ability which individuals possess to varying degrees. Reviewing my own audio-taped think-aloud protocols at some years' distance, I was struck by how much of the content consisted of 'affective' comments rather than directly task-relevant cognition, and I noted several spans of protocol that I would be

reluctant even now to release unedited. Verbal report data contains a good deal of such affective material (Afflerbach 2000) which informants must, to the extent that they choose to do so, edit for release in real time. What gets thought-aloud cannot be taken back, and I have on several occasions encountered uneasiness among post-task retrospective interviewees about what they have revealed in protocols. (In such cases the ethical course is to offer to delete the informant's verbatim speech and substitute a paraphrase.) I return to the problem of low productivity in think-aloud in my discussion of 'annotated cloze' in chapter 7.

4.14 Pair and solo reporting conditions

Given that according to the Ericsson & Simon 1984/1993 model a think-aloud verbal report should essentially be a monologue, my decision to have informants report not only individually but also in pairs requires some justification. I focused initially on solo condition reports for two main reasons. Firstly and perhaps more importantly, solo condition reporting most closely reflects the condition under which real-world cloze tests are taken. Secondly, I hoped to maximise the number of protocols I could gather, in the hope of obtaining a sample large enough to allow at least basic statistical analyses. That said, I was aware of the apparent benefits (Haastrup 1987; 1991) of having informants report in pairs (or 'dyads' as Haastrup labels them.) These benefits may include not only a greater more reporting in terms of volume but also data which is more readily interpretable and which might contain insights not available from solo-condition protocols. There seemed, then, to be some value in obtaining pair-condition reports for purposes of comparison.

Although I was open to the idea of collecting data from paired-informants, two factors led me to begin doing so (in parallel, as it were, with solo-condition reporting) rather earlier than I had anticipated. The first of these was that a disappointingly high proportion of the early solo-condition protocols I gathered from GL1 informants proved not to be very illuminating about their processing of cloze deletions. The protocol data was often sparse, and often much more difficult to interpret than I had anticipated from the reports of other users of verbal report procedure. The second factor was in the German university system assignments appeared to be allocated to a much greater extent (than I recalled from education in the UK or even in Japan) to students working in pairs or small groups. This tradition had to some extent permeated the English courses from which the majority of my GL1 informants were drawn, and students there were well used to completing tasks in pairs or groups. It did not take long before informants (all of whom had experienced solo reporting in at least one orientation) began to ask whether they might report in pairs. Those who did so claimed (with few exceptions) that they found the experience of paired-reporting 'more natural' and 'less stressful' than solo-condition. Word of this quickly spread among my informant pool, with the result that a number of informants expressed a very firm preference for pair-condition reporting.

In chapter 5 I discuss the concrete products of solo- and pair-condition in more depth. To the objection that pair-reporting diverges from what is essentially the

think-aloud procedure—and might better be seen as a conversational 'talk-aloud', I would reply that it represents only one of a number of deviations from that model which can be found in the literature. Some of these apparent gaps between theory and practice are taken up below and in chapter 6.

4.15 Think-aloud in theory & in practice

There is widespread agreement (Afflerbach 2000) that the central event in the rehabilitation of verbal report data (of which think-aloud is a key form) in the social sciences came with the publication in 1984 and subsequent dissemination of the Ericsson & Simon model of how verbal reporting can be made to provide valid and reliable data. Although not all studies that make use of think-aloud mention a theoretical basis for the procedure, with few exceptions (Boren & Ramey 2000) those which do so cite the Ericsson & Simon model. Ericsson & Simon 1984/1993 (and particularly the original 1984 edition) can reasonably be seen as *the* authority-of-choice concerning verbal reports.

At the risk of over-simplifying, the basis of Ericsson & Simon's model is that verbal report can provide valid data insofar as it is taken to provide a picture of the items of information the reporter attended to in her short-term memory (STM) (alternatively, in 'working memory' cf. Baddeley 1986) and the order in which these items were heeded. Boren & Ramey 2000 compare this to "time-stamping" the contents of STM. Ericsson & Simon distinguish three levels of verbal report data, each less reliable (and/or less valid?) than the one before. This lessening of reliability is caused by the increasing amount of interference to the informant's

processing at each higher level/part-of-sentence:

Level 1 data, for Ericsson & Simon, consists of data which does not need to be rendered into verbal form before being reported. Examples cited include number sequencing in mathematical problem. On the face of it, data generated by the processing of language tasks would appear to belong to this level.

Level 2 data must be transformed into speech before reporting, and in concurrent think-aloud this must happen in 'real time' alongside other aspects of the task. Level 2 data is also seen as reliable provided no other cognitive processes were applied to it beyond that of rendering it into speech. Stimuli for level 2 data would include visual images and abstract ideas.

Level 3 data in Ericsson & Simon's model are those verbalizations which entail cognitive processing beyond that involved in rendering them 'speakable'. It is worth quoting here from Ericsson & Simon (op.cit.:79–80):

"A third level of verbalization requires the subject to explain his thought processes or thoughts. An explanation of thoughts, ideas or hypotheses or their motives is not simply a recoding of information already present in STM, but requires linking this information to earlier thoughts and information attended to previously."

Level 3 data is inherently unreliable in Ericsson & Simon's view, and they suggest that it is this level of data against whose use Nisbett & Wilson 1977 argued. Other verbal report content excluded from consideration (Ericsson &

Simon, op.cit,: 223) would include "feelings" (which term these authors give no reason to think would not include at least partly affective verbalizations such as perceptions of difficulty.) Boren & Ramey 2000:262 further note that:

> "[u]nlike [classical] introspection and other techniques which value the content of verbalizations, Ericsson & Simon's model values 'hard' (attentive, sequential) verbal data [...] under no circumstances are verbalizations to be valued for their subjective content."

Ericsson & Simon (1993:xxxi) themselves draw attention to the "great diversity in instructions and procedures used to elicit concurrent and retrospective verbalizations." Although, they say, "many so-called think-aloud studies include elements [...] that would influence the sequence of thoughts" (ibid.) the effects of these elements may be less than that of most other experimental variables. Moreover, Ericsson & Simon's insistence that only data present in STM (and cognitively mediated only to the extent of being rendered, where necessary, into verbal form) can be seen as reliably reportable is, as Afflerbach 2000:172 makes clear, suspect. Although in theory the content of STM and that of long(er)-term memory are clearly distinguished, says Afflerbach:

> "in real time differences blur. An online and concurrent verbal report can dodge in and out of retrospection based on the length of the verbalization, the instructions to subjects, and the nature of the task."

Another key tenet of the Ericsson & Simon model thus begins to appear shaky.

The implications of these above for the application of think-aloud procedures to events such as cloze processing offer food for thought. Although Ericsson &

Simon 1984/1993 devote scant attention to text-based tasks, researchers investigating language tasks have felt free to apply think-aloud and to justify this in terms of the Ericsson & Simon model. A close reading of Ericsson & Simon gave pause for thought, but I chose in the end to swallow the doubts I had and to gather verbal report data about cloze recovery. I take up the wisdom of that decision in chapter 8, where I illustrate my contentions that (1) 'natural' think-aloud data quite commonly contains verbalizations at all the 'levels of data' posited by Ericsson & Simon and (2) if we accept only verbalizations that this model regards as reliable, we might gain even more limited insight at all into what goes on in cloze test-takers' minds.

Arguably the most explicit critique of the, as it were, factory-floor application of think-aloud methodology can be found in Boren & Ramey 2000 . The main thrust of their critique of think-aloud is not so much to question the Ericsson & Simon model itself as to point out that "theory and practice are out of sync", that is, researchers are using the model to justify things it does not sanction. The authors cite many instances of studies (mainly in the field of usability testing, an area which may have something in common with test construction) which incorporate or even focus on 'Level 3' material which is specifically excluded under Ericsson & Simon's scheme: explanations and evaluations of actions or event, emotional responses to these, etc. They note that researchers may interrupt informants during the task to seek elaborations on aspects of the verbal report, thus violating the monologue assumption in the model as well as the stricture against eliciting

subjective content, and in many other ways deviate from what is supposed to be the theoretical rationale for their data-gathering. As Afflerbach 2000:165 puts it:, although theoretical challenges to think-aloud have been made to think-aloud (and many answered), "in practice they receive intermittent attention".

In the application of verbal report to the study of language tasks perhaps the most explicit attention is paid to the putative mismatch between model and everyday practice by Pressley & Afflerbach 1995. These writers survey some thirty-eight separate verbal report studies of first-language reading, and the thrust of their argument may be less that any given model must be adhered to than that a common language of verbal reporting is needed if studies are to be mutually interpretable and comparable. Although all the authors in Faerch & Kasper 1987 made use of verbal report techniques of one kind or another, even at that time some appear to have felt uncertainty about aspects of the procedure: Haastrup 1987 (in that volume) suggests that methodological refinements are needed to verbal report, specifically with reference to second language acquisition studies. Waern 1988 and Smagorinsky 1998 both appear to take a position of 'critical support' for think-aloud, while Cohen 1997 clearly acknowledges the need for improvements in the application of verbal report procedures, and sub-classifies verbal report elicitation procedures according to the immediacy and directness with which they gather data. For Cohen, as for Ericsson & Simon, the faster and less mediated the better.

Among more recent 'handbooks', Green 1998 does not raise questions about how closely the theoretical basis and derived practice jibe, except to note in that "[o]ne of the problems in introducing a new methodology [...] is that researchers may quickly adopt the general ideas but forget to pay sufficient attention to the specific details" (op.cit.:119.) Alderson 2000:334 notes that thinking-aloud may not be the universal skill Ericsson & Simon claim. At the level of individual verbal report-based studies, it is not hard to find examples that incorporate elements which Ericsson & Simon might label inappropriate to think-aloud: Yi'an 1995, for instance, looks at how test-takers answer 'Why?' questions, and Haastrup 1991 has her informants report in pairs, i.e. in a dialogue rather than the monologue the model specifies.

To sum up, it may be reasonable to posit among writers on verbal report in the field of language study a rough consensus that the procedure is in need of refinement and/or overhaul, but the information it provides may generate insights unobtainable by other methods. This position implicitly acknowledges that verbal report is, in the words of Pressley & Afflerbach 1995: "a maturing methodology with much interesting work already accomplished and considerable work to be done".

4.16 Instructions & Subject-training

In this section I look at the question of training informants in the data-collection task, and at some of the pros and cons of such a step. I go on to discuss a compromise procedure which I believe helps reduce the risk of biasing

informant's task processing while still providing a degree of guidance.

Sources of variation

Ericsson & Simon claim (1993:224) that : "Thinking aloud [...] can be elicited almost instantaneously by the appropriate instruction from virtually all human adults." They also note (1993:xxxi) that the instructions used to elicit concurrent verbalizations, i.e. think-aloud, vary greatly across studies. Given that most think-aloud studies have managed to extract at least some data from at least some of their informants, the first statement (if true, which it is not) would imply that there must be a undefined *set* of possible appropriate instructions.

In fact, individual informants can vary markedly (Gilhooly 1986) in the degree to which they verbalise their cognitive operations. Variation also exist in the kinds of operations informants report. This variation is presumed (Grotjahn 1987; Neumann 1995) to be, to some extent at least, a function of the stimulus task set and of the data-collection procedure(s) employed. There is, however, also conspicuous variation in informants' productivity even within the same data-collection setting. Neumann 1995:98ff compares two studies (Zimmerman rev.ed. publication date 2000, and Krings 1992) of L2 writing processes which used broadly comparable subject groups and stimulus tasks, and notes the sparseness of metalinguistic (in the sense of 'global' structuring of the text under construction, i.e. what other writers might label metacognitive) utterances among Zimmerman's subjects, and the much higher production of these by Krings' informants.

The major difference Neumann identifies between the two studies cited is Krings' use, and Zimmerman's non-use, of subject training, and in her view this is the source of much of the divergence between the verbalizations of the two groups of subjects. The metalinguistic utterances found in Krings' data are, Neumann suggests, "not to be expected...from untrained informants." Although the verbal reports of Krings' subjects are both more informative about 'high level' writing processes and more explicit overall, Neumann questions whether these high level verbalizations "...actually [provide] an authentic insight into the writing process, as they are based on a higher level of abstraction than the think-aloud itself." Again, we see the tension between elicitation of what appears to be informative data and the constraints (more abstract equals less reliable) of the model. It should be added here that Neumann (op.cit.) questions whether studies involving trained informants are even comparable with involving untrained informants (cf. Pressley & Afflerbach 1995.)

Options in subject-training

A possible corollary to Ericsson & Simon's observation, above, about the power of the appropriate instruction is that training in the task may not be necessary. My reading of Ericsson & Simon 1984 suggests that the above would be a fair reflection of their position then, which seemed to be wary of the risk that training would bias the resulting think aloud data. The 1993 revision of their book (consisting mainly in the addition of a preface) explicitly allows a role for practice sessions (followed if necessary by repetition of the instruction–practice cycle) but

the authors do not address the issue of training in any real depth. We are told (:82), without further comment, that in some studies, "more extensive warm-up procedures are used explicitly to *train* [italics in original] the subjects to conform to the [think aloud] instructions." The training used in think aloud is, however (ibid.) "negligible compared with that employed in classical introspection experiements." 'Warm up' and 'training', for Ericsson & Simon, appear to be practically synonymous.

Many studies in the think-aloud literature at least touch on the question of whether or not subjects should be trained in the task of verbalizing their on-task operations. Haastrup 1987, 1991 includes her written instructions to informants in 'pair' think aloud lexical inference tasks, but does not appear to have employed warm-up of training sessions. Block 1986 mentions, but fails to describe, a 'short orientation' and her 1992 paper mentions (but *ditto*) a 'training session' prior to the think aloud task. Cavalcanti 1987 suggests that, left to their own devices, untrained informants tend to read aloud fairly long spans of text about which they then–in effect–retrospect. This particular behaviour, she argues, can be obviated through training.

Risks of training

A degree of caution is called for here. My own experience suggests that the above is just one among many behaviours observable in think aloud, and may appear in a protocol whether or not the informant has received instruction in think aloud procedure. It appears to be quite common, moreover, in translation tasks where no

think aloud requirement is set; as we shall see, translation of spans of text features heavily in many informants' think aloud protocols. It may indeed, be possible to obviate the behaviour by specifically instructing informants *not* to verbalise in this manner and/or correcting them when they do, but what if this is in fact an *authentic*, natural processing behaviour? During my self-as-subject explorations of think-aloud, I found myself reading aloud spans of German text whose meaning I could not immediately grasp. During and immediately following that reading-aloud I would perform a kind of word-by-word or chunk-by-chunk translation-cum-interpretation. I still go through the same processes (silently, I hope) when reading in other languages, which suggests that read-and-retrospect/interpret may be a fairly common way of processing difficult text. My point here is that, unless a given process or operation can be shown to be no more than an artifact of the task or data-collection condition, we should think twice before attempting to 'train it away'. Except–perhaps–when replicating an earlier study, we cannot know which on-task behaviours are authentic until we have had experience of eliciting data via the chosen procedure. What I take to be Ericsson & Simon's continuing wariness about training may be well-advised.

4.17 Modelling of task behaviour

'Warm-up', to me at least, connotes a largely unguided exposure to a task which the user or performer already understands, and which allows some degree of practice or rehearsal. It is uncontroversial that practice in a task typically leads to faster, better, or more efficient performance. The introduction of a behavioural model, however, may introduce new behaviours or modify existing ones.

That informants may tend to emulate modeled verbal report behaviours is suggested by the trial protocols of six GL1 informants who were exposed to such models. Three subjects listened (independently) to one of two different audiotape models of verbal reporting while looking at the stimulus cloze passage. On the first audiotape model I verbally reported my operations in filling blanks on a short German-language cloze passage prepared by a helper, and in doing so I attempted to expand on or emphasise 'within text' passage cues while 'downplaying' cues based on extratextual knowledge. In another audiotape model I repeated the procedure with the same German-language cloze passage, but this time I attempted to expand on extratextual cues while downplaying 'text internal' cues. 'Playing up' and 'playing down' was realised by talking at greater, or shorter, length about a cue, and/or by 'flagging' it for relevance with markers such as "Aha!"; "Moment mal..." ('Just a second...'); "Also.. hier steht's..." ('Well, here it says...')

It is clearly very difficult to quantify the above emphases, or the intuitive observed bias in subsequent subject protocols. My impression, however, was that all three subjects exposed to the model emphasising extratextual knowledge cues appeared to echo this emphasis in their protocols, while those exposed to the 'internal' model seemed to restrict themselves more to textual cues. This impression was echoed by two consultants, one a native-speaker of German and the other of English. These were shown the cloze passage and (with the permission of the informants) auditing the six protocols 'blind', they were asked

to allocate each to one of two categories:

(1) 'Uses information from within the passage more' or

(2) 'Uses information from outwith the passage more'.

Both consultants allocated the protocols to the same categories as I had done, noting only one instance of serious doubt. To have any value this study would have to be repeated with (a) larger samples and (b) more precise quantification of both task difficulty and subject behaviour. It would not be surprising, however, if effects of modelling/training were to prove significant in conditioning informants' verbal reporting. After all, most informants have little or no previous formal experience of this type of task—although, as Afflerbach 2000 suggests, verbal report may be close to the Vygotskian notion of "inner speech") and might thus be expected to resemble that behaviour.

4.18 Another experiment

In a subsequent study, I gave a group of twelve 'naïve' Japanese L1 informants the usual instructions in verbal reporting, after which half of the group (randomly allocated) were taken to another language laboratory room to begin their verbal report session. The remaining half then listened to extracts from the audiotape protocol of a Japanese L1 informant who had made use of a particular cloze processing strategy which I have labelled 'skipping.' In this mode the informant read the entire text (in large part aloud) from start to finish, at fairly high speed, inserting fillers where she could, and where she could not recover a blank she more or less immediately carried on reading as before. Although other informants

had been heard to process parts of a text in a similar way, as an overall, consistent strategy this was statistically highly unusual. Out of the six informants who had received the usual instructions in verbal reporting *plus* subsequent exposure to this informant's protocol, however, three appeared to consistently apply the same strategy, while a fourth used it quite extensively in the first half of the task. (One of the remaining two informants in this group verbalised very little.) Of the other group, only two informants made some use of 'skipping' in the early stages, but none used it anything like as consistently as those in the 'model' group. This result suggests that models of verbal report may quite strongly affect informant behaviour.

Numerous writers, in fact, have expressed awareness of the risk of bias that subject-training brings. Gerloff 1987 suggests that offering informants training in verbal reporting risks making informants "look more alike" in their processing. Neumann 1995 suggests that different cognitive processes can be ranked along a "scale of consciousness" (*Bewußtsheitsskala*) so that individual processes are more or less accessible to conscious attention—a precondition for verbalization. While training may raise a particular process or processes on the scale, this clearly distances the result from a 'natural' task condition. Grotjahn 1987 offered his C-test-taker informants an audio recording "illustrating thinking aloud" and made clear to them (in a slightly ambiguous rendering into English) "the exemplary nature of this recording."

4.19 No overt model of appropriate task-behaviour, but practice

On the basis of the above findings and arguments, I chose not to provide my informants with a model of verbal reporting. Instead, I limited pre-task treatment to a fixed set of written instructions , combined with a short (10—15 minute) individual practice cloze task (see below). This practice session was loosely comparable with the 'trial' sessions used by Ericsson & Simon 1987 and Krings 1989, with the added (and in the event only partly realised) intention that it would allow me to remove from the informant pool any who had produced little or no verbalization. In the practice session the instructions and requirements of the think aloud procedure were explained to informants (either individually or in a group, and where possible in a LL-equipped classroom) and then each informant was presented with a short practice think aloud task. Her processing of the practice task passage was observed as closely as possible, but my interventions were limited to (1) encouragement to "Keep talking" and (2) answering procedural questions ("Do I need to put in a word here?"; "Is it really always only one word?", etc.) Individual practice was felt to help reduce the risk that one informant's behaviours (cf. Gerloff, op. cit.) would transfer to other subjects, which my own observations of classroom tasks had suggested might occur where one informant was perceived by another nearby as being linguistically more able.

Practical advantages of the practice session

As well as minimizing the risk of 'cross-contamination', the introduction of a practice session may have two other important practical benefits for the data-collection process. Firstly, observation suggests that short trial tasks quite

spontaneously lead informants to verbalise more readily or 'fluently' as they work through them. Subjects are unlikely ever to have been set a verbal report task before, and so a practice effect is perhaps to be expected. Given that informants tend to verbalise more in the later stages of a task, it is preferable that any practice effect should occur before the actual data-elicitation session. The practice session also acclimatises subjects to the 'keep talking' instruction they are likely to be given (after *n* seconds' silence.)

Moreover, practice tasks allow informants to acquire a feel for the range of possible target items. In several instances, informants who had been informed both orally and in writing (and appeared at the time to have understood) that only one word had been deleted at each blank; that the deleted item could belong to any word class; and that a word had to be inserted into *every* blank seemed to internalise some or all of these points only when faced with actual recovery. Obviously, an informant's failure to grasp the ground rules of the task in hand reduces the validity of any data gathered, and it is vital to minimise the risk of misconceptions persisting into the actual elicitation sessions.

Avoiding bias from the researcher

I was concerned, in trial or practice sessions, to avoid introducing non-spontaneous or inauthentic behaviours via my own instructions or explanations (cf. chapter 8 for a brief discussion of whether I was successful in this.) Interruption with the instruction to "Keep talking" is standard in the field, and as it is intended only to maximise the volume of verbal reporting it is held

(Ericsson & Simon 1984/1993; Block 1986) to be acceptable. Care was taken that informants should not, however, receive any guidance as to what they should be verbalizing about. Those who sought approval were given it: they were told that what they were saying was interesting and useful, and invited to say even more. No practice retrospective interview was given, however. This was due to (a) the fear that hearing the kinds of questions I might ask would bias the informant's task-processing in unpredictable ways, and (b) the fact that until I listened to informants' protocols I had little idea of the questions I would put to them.

To sum up, in the present state of knowledge about the range of 'authentic' on-line behaviours open to informants, I was wary of limiting or distorting these through task-training. Carefully controlled practice sessions, however, appeared to offer clear benefits in terms of the volume of data likely to be collected in the data-elicitation task itself, and (except for the few who realise that they cannot successfully think aloud) in terms of informant confidence.

4.20 The physical conditions of data-elicitation

The (abbreviated, 32 deletion) OLYMPICS task passage as used in data gathering is shown in Appendix 1, along with the English version of the task instructions which accompanied it. The physical situations in which I was able to gather think aloud data varied from language laboratory rooms (in which the LL equipment might or might not be functioning properly), through regular classrooms, to my own slightly cramped office. Only solo condition think aloud data could effectively be gathered in a LL setting, but in my own office and in other rooms it

was possible to accommodate solo- or pair-condition sessions. It was necessary, in a small office, to be fairly discreet about the observations I made of informants at work, but the relative lack of distraction allowed me to focus on the informant or pair at hand. A variety of cassette recording equipment was employed, often with two recorders running at the same time as one set-up was trialed or used as a backup. Both 'clip-on' and tabletop microphones were employed, but the latter were found to produce better quality recordings. Due to equipment limitations all recordings were monaural. Recordings were made onto C90 cassettes to provide 45 minutes recording time per track.

Solo- & Pair-condition reporting

The notion of having informants report in pairs rather than individually was broached in section 4.14, and although it was never intended to be my primary data-elicitation mode, the fact is that two-thirds of my think aloud informants from both language backgrounds have reported in pair-condition: my use of it thus requires some justification here. Although researchers like Haastrup 1991:85 make a conscious decision to prefer pair-condition think aloud (here, PCTA) over solo-condition (here, SCTA), on the basis that:

> "[…] by using pairs ['dyads'] one stimulates informants to verbalise all their conscious thought processes because they need to explain and justify their hypotheses […] to their fellow informant."

adding that, in the study she describes, it is unlikely that a solo-condition reporting format would have produced protocols as informative as those obtained via pair reporting. The use of the phrase "verbalise all their conscious thought

processes" suggests that, for Haastrup, pair reporting extends the elicited dataset along the axis of breadth as well as depth, if this distinction is meaningful here.

There are, however, grounds on which to question the use of paired think aloud. One caveat about PCTA as applied to something like cloze is that it diverges from the essentially solitary nature of test-taking. (Haastrup 1991 feels that lexical inferencing is likely to be a shared task under everyday circumstances, so that PCTA is quite authentic.) This aspect of the procedure was explicitly commented on by a number of my informants. In the words of one "I think my partner's suggestions I was helped a lot. I wouldn't have got so many [cloze fillers] if I [had been doing the task] alone." This remark reveals the bias shown by the perhaps inevitable concern felt by many JL1 informants with 'cloze task done how well', rather than the research focus on 'cloze task done how'.

Another possible objection is immediately apparent: it is unfeasible for more than one participant to verbalise at any given moment. Assuming that the partner who does not have the floor does not suspend her thought processes entirely, these are lost to concurrent reporting at least. Indeed, pair reporters may anyway verbalise only a subset of their thoughts. The Gricean conversational maxim of economy summed up in "Do not tell your partner what you think she already knows" surely applies to pair verbal reporting as much as to any other conversation. (This maxim may be assumed to apply less strictly in 'tutor' condition pair reporting, in which one participant is (or pretends to be) markedly less able at the task, and thus has to be instructed by her partner in how to go about processing the task in hand.)

Yet another reason to question PCTA procedure is that it inevitably introduces a range of almost certainly complicating, if not biasing interpersonal factors. One such factor (as Haastrup 1991:84ff notes) is the degree of cooperation and consensus between partners. On several occasions conflict between GL1 informants became quite serious, and though in most cases resolved, in two it led to a sudden and premature end to the paired data-gathering session (whose protocols had to be discarded.) Conflict between Japanese L1 PCTA informants was never clearly obvious to me during the task itself—although a Japanese colleague who partially observed one such session in progress claimed to detect signs of mild discord between the partners—but reports of disagreements did surface in several individual post-task interviews with JL1 informants.

The (im)balance of power in a PCTA task may also affect what gets externalised (but cf. a report in chapter 6 which suggests that this may not be a major issue.) Despite my efforts to match paired GL1 informants according to measured TL proficiency and 'fluency' (by which I mean less the extent or accuracy of their TL language ability than their willingness to interact with a partner and with the task) some imbalances did arise. To the extent that one informant might feel she was being 'talked down to', or felt unable to play an equal role in the task, this could give rise to unease. Some of the 'stronger' informants in such situations also remarked, post-task, on their discomfort. The above applies only to researcher-selected pairings, which in my own case were too often fortuitous and depended on who turned up at a given time. Self-selected pairs (the majority in

my samples, as ability to select a partner with whom she felt comfortable seemed to be an important factor in whether or not a potential informant volunteered) seemed better able to cope with disparities in ability. Most pairs self-selected on the basis of background and/or friendship (and, yes, perhaps also according to perceived TL ability) and although it complicated the task of creating other pairs, I felt that this had to be accepted.

The above factors contribute to an issue far from trivial to the transcriber of PCTA recordings. These can be very long, and on average, those I gathered were some 50% longer than their solo-condition counterparts. Solo-condition recordings may (despite reminders to 'Keep talking') contain a good deal of silence whereas pair-condition recordings contain far less. Moreover, a good deal of PCTA protocols may consist of material which is not strictly or directly task-related, and even that content which is clearly pertinent to the task may be expressed at much greater length. Compare these extracts from a solo JL1 informant and pair-condition informants working on the same cloze item:

SCTA JL1 informant Akemi

"[...] the exact of events is uncertain.. but events.. include?..included ISN'T IT? Hmmm.. such as.. FOR EXAMPLE RIGHT?..hmmhmm [5 secs] OH YEAH.. the exact number? number of events.. exact number.. OR IS IT TO DO WITH THE ORDER? (*junjo*).. EITHER ONE IS OKAY, RIGHT?.. I'LL CHOOSE NUMBER [...]"

PCTA JL1 informants Mayumi and Yumiko:

Mayumi "exact number PERHAPS?"

Yumiko "hmmm.."

Mayumi "WHAT D'YOU THINK?"

Yumiko "THAT COULD BE RIGHT.. I CAN'T THINK.."

Ma. "IF YOU HAVE ANOTHER IDEA.."

Yumiko "NO.. BUT.."

Mayumi "WHAT?"

Yumiko "DOES schedule FIT HERE?"

Mayumi "schedule?"

Yumiko "schedule.."

Mayumi [indicates uncertainty/puzzlement] "YOU MEAN LIKE A LIST?"

Yumiko "LIKE A TIMETABLE.. LIKE ON A SPORTSDAY.. YOU KNOW?"

Mayumi [laughs, mimics chanting spectators] "YAY! HANG IN THERE! DON'T GIVE UP!"

Yumiko [laughs] "YEAH YEAH.. JUST LIKE THAT.."

Mayumi "WELL THEN.. the exact timetable? SOUNDS A LITTLE STRANGE, NO?.."

Yumiko "the exact timetable of events is uncertain.. IT'S NOT CLEAR?"

Mayumi "THAT'S RIGHT.."

Yumiko"WELL THEN.. LET'S SAY list OR schedule.. THEN WE CAN GO ON.."

Mayumi "YOU'RE RIGHT.. list OR schedule?"

Yumiko "STICKS-STONE-PAPER?" [This is done.]

Mayumi "schedule THEN"

Yumiko "OKAY"

A final ground for concern about PCTA has to do with Haastrup's rationale for pair reporting cited above. Haastrup (1981:85) herself notes that, in a lexical inferencing task, a pair informant may suppress her guess about a word's meaning "for fear of being ridiculed." I never saw or heard an informant overtly mock another's suggestion except in a very few instances in which friends had self-selected as partners; in these exchanges a certain amount of teasing seemed to be expected, and no offence appeared to be taken. A wider problem, however, may stem from the fact that (as observation of PCTA and protocol review confirm) paired think-aloud follows in large part the structures and strictures of

regular peer conversation. Partners thus have to negotiate turn-taking, and disagreements and counter suggestions have to be presented in such a way as to avoid giving offence. Something very like the Gricean maxim of economy seems to apply here, moreover, in that informants may shy away from telling their partner what they expect she already knows. A number of individual post task interviews (only some of which were audio-recorded, mainly due to lack of equipment) produced comments such as that below, in which a GL1 informant explains why she failed to mention in the PCTA session an item of information (the name of a specific event within the pentathlon) which she volunteered in the post task interview and claimed to have heeded during the task:

Researcher: "so you read that somewhere? that the longjump was part of it.."
Informant (Kayuko): "yeah.. [I] don't know exactly when or where.. I'm pretty sure though"
Researcher: "but you didn't.. [Ka. anticipates remainder of comment and rejoins] (mention that to your partner)
Kayuko: "no.. I thought.. it would be clear.. that she would know it too"
Researcher: "so you were happy [just to fill] with jumping?"
Kayuko: "yeah.. APART FROM THAT I'M NOT SURE IF YOU CAN SAY longjumping OR just longjump"

In addition to outright avoidance of information presumed to be shared, GL1 protocols reveal numerous instances in which an informant introduces information she feels may be known to her partner with L1 qualifiers such as: *'Klar'; 'Natürlich'; 'Logo.'* These are functionally analogous to English 'of course', 'stands to reason' etc., and I interpret them as conveying something along the lines of "I don't need to tell you this, but..." In other words, GL1 paired

informants often appear to be trying *not* to 'talk down to' a partner; the Japanese language has many, many analogous gambits. Informants' reluctance to raise information or insights that they feel their partner may already be aware of is clearly undesirable to the extent that it constrains the reporting of the text information or other knowledge that informants have attended to in recovering deletions.

Should one partner disobey the 'economy maxim', however, conflict can ensue. In my observation of PCTA sessions I noted a few instances in which one partner appeared to react negatively to (what at times appeared even to the observer as) the other's overbearing behaviour. The offended partner might fail to react in any obvious way to the offending behaviour, or might become less active in the task in an apparent withdrawal of cooperation or goodwill. Ironic comments were also heard 'oh thank you.. I'd never have though of that!' as well as more explicit indications that the information was already known to the speaker. It was in the rare instances in which the offending partner failed to note these hints, and modify her behaviour accordingly, that conflict could escalate. Once again, however, it must be emphasised that genuinely self-selecting pairs appear to be less prone to discord.

Haastrup's 1991 informants appear to have found the experience of pair-reporting enjoyable. For at least some of mine, it seems almost to have been a precondition for participation. For all its disadvantages of protocol length and risk of complicating interpersonal dynamics, without exception, those JL1 informants

who had experienced both solo- and pair-reporting in practice-cum-orientation sessions (A few informants attended more than one orientation session before deciding to participate as informants.) told me that they found pair reporting much less stressful. Informants' affective response to the task is not (and I return to this point in chapter 8) something we should overlook.

4.21 Post-task interviews

Having outlined some aspects of the think aloud stage, I turn now to the role of post-task retrospection, and its use as a means of eliciting triangulatory or complementary data. Triangulation can be defined for our purposes here as "the use of two or more methods of data collection in the study of some aspect of human behaviour" (Cohen & Manion 1994:233) and is (ibid.) a technique widely subscribed to in theory but less often used in practice. The most common source of triangulatory data in language task-oriented verbal reporting lies in post-task retrospections or interviews, and I discuss below some positions writers and researchers have taken up around this topic.

The problem (or not) of timing

One concern about retrospective reporting has traditionally lain in its distance in time from the original cognitive event. In fact, Ericsson & Simon 1993 distance themselves from their "earlier caution" (op.cit.:xvi) about retrospective verbal report in the first (1983) edition of their book, and indeed for a circumscribed set of events of short duration (none, as far as I can tell, of the sort studied by language researchers) they recommend that retrospective data be preferred to

concurrent. They discuss at length (1994:xx ff.) studies that elicited retrospective data, but without making any serious methodological criticisms of these. At other points Ericsson & Simon seem to view retrospection entirely positively, proposing "[the substitution of] retrospective for concurrent reports" (1994:253) in order to increase the amount of verbal data which can be collected. That said, Ericsson & Simon seem, overall, to persist in their original concern that delayed verbal reports run a greater risk of replacing a true picture of what occurred in the reporter's mind with a (partially) reconstructed and thus unreliable one.

There are, however, grounds to doubt (Afflerbach 2000:172) that the neat dichotomy of concurrent report and retrospection is a valid one in practice. The caution expressed by Ericsson & Simon 1983 does not, anyway, appear to have been much of a barrier to the elicitation of retrospective reports in relation to language tasks, and the potential value of retrospection in supporting, elaborating (and even challenging?) concurrent data has attracted many researchers such as Haastrup 1987, 1991; Ridley 1997, and the authors of many of the studies collected in Faerch & Kasper 1987 as well as those surveyed by Pressley & Afflerbach 1995. Haastrup 1991:86 explains her use of retrospection as follows:

> "The main purpose of the retrospective session is to make the verbal reports more complete and to make the interpretation of the reports more reliable by asking informants to make more explicit what they had hinted at earlier."

This quotation sums up quite succinctly the problems of concurrent think aloud which I will illustrate in an upcoming chapter: think aloud data tends to be

incomplete and hard to interpret, and so we have to ask the informant to help. Perhaps open to question in theory, the elicitation (post-task) of retrospective data may be indispensable in practice.

Options in post-task interviewing

To some extent the form and content of a post-task interview (or PTI) will be defined by the task it focuses on and the participants involved. Some uncontroversial guidelines may be set out, however. Firstly, the PTI should take place as soon as possible following the task itself. It must be borne in mind, however, that many think aloud informants find the task fatiguing and even stressful. (Around half of my own informants have requested a 15-20 minute break between task and PTI, and a few have even changed their minds after the task and declined to be interviewed at all.) Although Gass & Mackey 2000:99 merely ask that researchers "[c]onsider carefully the implications" of using the informant's L1 or the TL as the medium of retrospection, I take it as axiomatic that the same strictures apply post-task as on-task: the informant must be given an authentic choice of which language to use, and it must be clear that she can switch languages as she feels necessary or desirable.

Secondly, some form of memory support may have to be made available. There is, as Afflerbach 2000:170 points out, "often a startling lack of detail provided" in published reports of verbal report data-elicitation sessions, and discussion of memory support is one area which quite often appears to slip through the cracks. Memory supports, which in think aloud studies will typically tape the form of the

audio-recording the informant made (perhaps alongside a copy of the task) may be used in a variety of ways and with a range of functions:

(1) The tape may be played and the informant asked to comment on its content.

(2) The informant may be asked to comment on her processing of a given span or item and the relevant part(s) of the recording then played

(3) The informant may be allowed to listen to her own recording, either (3a) in whole or (3b) in part, and under her (3c) control or (3d) that of the interviewer) and asked to elaborate either directly or onto a second tape.

I have trialled almost all of these formats, and can offer the following observations. Option (1), above, is, as far as I can tell, the norm in studies that employ PTI. Provided the interviewer has noted the parts of the recording or task in which she is most interested, these may be accessed fairly quickly. Informants may or may not be able to elaborate on or clarify the tape content, but (unsurprisingly) any comments they do make are likely to be in line with what was heard.

Option (2) is more problematic in that the informant's retrospections may be at odds with what she subsequently hears on the tape, or (perhaps more commonly) evidence for some part of her retrospection is simply not there on the recording. This can lead to feelings of discouragement or, if she feels that she has been 'set up' by the interviewer, to a reluctance to continue.

I trialled option (3) with a number of JL1 informants, and in the hope of reducing the informants' stress and my own I used a 'take-home' format (i.e. (3a)+(3c),

above) in which informants each took away a Xerox copy of their original task sheet, a dubbed cassette (or digital medium, for those who did not own a cassette player) copy of their recording, and a blank recording medium, along with the instruction to listen, as soon as possible, to their original recording and add any information or clarification they thought pertinent. The clear advantages of this format are, firstly, that the interviewer does not need to be present, so that the full informant group can be 'interviewed' without fatigue and/or (see below) a long waiting period. (If the think aloud task took place in a LL, the researcher may have been able to ascertain which informants' verbal reports require PTI 'in person', and may ask the others to take a short break and then review and expand on their recordings in the LL itself, if scheduling allows.) A second advantage of take-away retrospection is that the informants may do the task in their own time.

The disadvantage is that the retrospection is uncontrolled (or, rather, controlled only by the content of the original recording, and perhaps to some extent by the instruction to elaborate or expand on what has already been said.) and the resulting elaborations run the gamut from highly illuminating to barely informative at all. The informant may, in fact, elaborate at length on points which were already fairly clear, and/or omit to clarify those which were not. The risk of bias through leading questions is of course absent in unguided retrospection, but if this is not felt to be a grave danger then some written questions may be added to guide the informant in what to focus on in. The take-home task format works most efficiently if the researcher/interviewer is able to identify points of interest or

concern to the informant before she carries out her review, and this may be no more than a list of individual items about which one would like to know more. (In the section devoted to NCR I outline what I have found to be optimal recording and review procedures.) As with all other retrospective tasks, it is better to conduct them as soon as possible after the task-session.

Group retrospection

Alderson 2000:333 notes that 'introspection' may take place in groups, a simultaneously more (more people) and less (each has fewer chances to speak) efficient variant on Haastrup's (1987; 1991) dyad reporting and recall. This idea may have possibilities even in cultures in which consensus is valued, for despite the self-applied 'group-minded' image of Japanese society, I have found (Gibson 2002) that younger Japanese informants at least appear to report opinions and reactions quite honestly in a group setting, and that dissent is not the taboo it may once have been. I have only conducted one trial of a focus group (n=4) retrospection about cloze, but I found the procedure reliable insofar as no processing operation or item of information was mentioned in the group discussion that did not also occur somewhere in the think aloud protocols. The Gricean economy maxim appears to apply to group conversations as well as to informant-pairs, however (see chapter 6), so that many 'obvious' operations (e.g. reference to syntactic knowledge) and information points went completely unmentioned by the group participants. Omissions like these might be minimised by a special instruction to 'speak even the obvious', and are to some extent offset

by the fact that one informant's comment may stimulate another to make an observation she might otherwise not have volunteered. Group retrospection or 'debriefing' about think aloud tasks is a potentially fruitful means of eliciting post-task data more quickly and efficiently, with the consequently lower risk (in terms of Ericsson & Simon's model) of forgetting or reconstruction, than might be possible with one-on-one interviews.

Bias from the researcher?

As well as the time-gap between events and retrospection, there has been concern (Afflerbach 2000) about the danger of stemming from the interviewer's approach. In face-to-face interviews, some attempt should be made to standardise prompts. It is thought (Ericsson & Simon 1984/1993; Cohen 1997) that prompts such as "Go on"; "Can you tell me more?" carry the least risk of bias, but specific questions may have to be put if the informant does not volunteer the information wanted. As Haastrup 1991:87 notes, it is difficult but necessary to strike a balance in the PTI between interviewer control of the material discussed and informant self-initiation. In my own conduct of face-to-face PTIs, I restrict myself as far as possible to the following prompts in the TL or in the informant's L1 as appropriate: "Go on"; "Tell me more about that."; "Did you do/notice anything else?"; "Do you remember doing [action mentioned by informant]/using [that] for any other blanks?"; "Did you know that before[hand]?"

Gass & Mackey 2000 discuss post-task retrospections under the label of 'stimulated recall method', and note that: "[t]here is often a relatively high level

of interpretation [required] in relation to data acquired through stimulated recall" (a caveat that applies also to concurrent think aloud data; see chapter 6) and note too that the data is often gathered, transcribed and analyzed by the same person. The risk of bias, then, is not confined to the actual post-task interview itself, but rather permeates the data gathered via retrospection. In short, not only may the informant retrospect inaccurately, but her retrospections may be wrongly interpreted by the researcher.

As the data discussed in chapter 6 will make clear, even the 'memory support' provided by replaying of their concurrent audio protocol is not always enough to allow an informant to recall her processing operation(s) at that moment with confidence. My position is that if the informant herself is not sure how to interpret a span of protocol data, the temptation to interpret it for her is best resisted. Qualifiers such as "X appears to..." should thus be taken as a best guess on my part, while "X thought she had..." represents the informant's own tentative recollection of her processing.

4.22 Further information from native-speaker consultants

Although I was by this stage moderately satisfied with my arrangements for data-collection, the onset of the German university examination and seminar-paper deadline period forced me (or, rather, my informants) to postpone several data-gathering sessions. I used the time to try to glean from native-speaker consultants what I hoped would be valuable information about two aspects of informant's processing behaviour. These were (1) what lexical, grammatical and

content knowledge the test-takers/informants could be expected to have acquired in the course of their language education and (2) their likely orientation or approach to the test-task. Taken together, these two knowledge-sources might, I felt, allow me to predict the relative 'clozability' of individual deletions, to anticipate the strategies and behaviours test-takers might use in filling deletions, and/or to anticipate problems they might encounter. My rationale for doing this was that it seemed useful to try to gauge the degree to which the kinds of people who might set cloze tests or use their results could judge how and/or how well test-takers might deal with these. This ability has been questioned by Alderson (cf. chapter 3) and was at that time a topic of informal debate among language teachers in the Berlin-Brandenburg area. The data presented below may be compared with that presented earlier in the chapter.

I discuss in chapter 6 the challenge posed in cloze passages by extant passage content, which think aloud data reveals may be a factor in task difficulty over and above the 'built in' challenge posed by the mutilation of the text. What I report here are the comments of 8 German and 5 Japanese native-speaker consultants whom I invited to predict the kinds of information test-takers might have access to, and difficulties they might face in processing the cloze passage used to elicit data in this study. I asked these consultants, all of whom were educated at least to high school level in their own countries' language education systems (and who were, like the informants themselves, predominantly female) to comment, either in writing (in their own time) or in a short face-to-face interview between myself and

one or two consultants at a time, on the mutilated version and its individual deletions. Consultants were mainly in their 30s. The German L1 group included graduate assistants (mainly intending teachers) and lecturers at two German universities. The Japanese group was slightly more heterogeneous, and included two university lecturers, two teachers at a *yobiko* college-preparation 'cram school' and a personnel manager. Their average age and level of education were comparable to those of the German group.

I used face-to-face interviews whenever possible for at least part of the session, and in these, consultants were first given the mutilated OLYMPICS passage. They were allowed to ask any questions they saw fit, and I confirmed their suggested fillers (accepting all semantically-acceptable candidates) or provided corrections as necessary. When two consultants were present, they tended to behave (although intermittently) very much as pair-conditions think-aloud informants do, discussing problem areas and suggesting and evaluating possible fillers. As consultants tended to be colleagues or other peers, or friends of mine, audio-recordings were not felt to be appropriate. Instead, I took notes of relevant information, and freely posed questions of my own.

Difficulty rating

One of the tasks I asked consultants to do was to rate the difficulty of each deletion on an incremental scale of 1 to 5. With the exception of difficulty level, I tried not to tie consultants down to 'quantitative' assessments. All, however, were apprised of my focus on the questions of how task-processors of the consultant's

own language and educational background might deal with a cloze task, and where any difficulties might lie. Consultants were also provided with a copy of the full, un-mutilated, passage and asked to comment on any difficulties which they thought undergraduate level informants might encounter, or any other points of interest or concern. I paraphrase or summarise below those comments and responses volunteered by at least two consultants from each language background. As not all consultants offered comment on all deletions, no useful average can be calculated. Instead, lowest, modal, and highest difficulty scale evaluations are given in that order.

German L1 consultants discuss the un-mutilated OLYMPICS passage.

Some general observations were noted and are paraphrased here:

> The topic [of the passage] should be familiar.
>
> Test takers may not know which sports were played in the ancient games.
>
> Overall the language is not too difficult for university level students.
>
> Some of these [blanks] are very easy, but some are difficult.
>
> In some cases it seems that no word would strictly be necessary.
>
> Sometimes more than one choice of word is possible. It might not be possible for most test-takers to choose the optimal word.
>
> The level of difficulty depends on how you grade the answers. If you allow a range of [fillers] with the same basic meaning then it becomes easier.

Consultants comments about individual deletions in the passage were also gathered, arte shown in figure 4.7, below.

Del.	Consultant comments (German L1)
1	It should be obvious that the missing word is a number. The English construction is similar to German '*alle x Jahre*'. Test-takers should know that the games took place every four years. The revived games adopted the same interval. 1,1,2
2	Test-takers may be confused by German '*eventuell*' meaning 'perhaps'. Something like 'lost' or 'abandoned' would fit here. ' The German word '*verlieren*' ('lose') could also be used in this context. Logically, you have to give up or stop being one thing in order to 'become' something else. 1,2,4
3	The logical stage between 'local' and 'international' is 'national'. 'Greek' may be chosen by some test-takers. 1,2,3
4	The same as German, 'rules against' You could use 'forbidding' or 'banning', but 'against' is most likely. 1,1,1
5	The equivalent German construction is close to the English. 1,1,2
6	German too would require the definite article. 1,1,1
7	'Official' makes it clear that records of some kind are being referred to here. 'Documents', 'records, 'plaques' could all be used here. Did the Greeks use paper at this time? 2,3,4
8	'Took place' is an English phrase which all should know. 1,1,2
9	Some may choose 'of' from the German '*von*'. Some preposition of location like 'below', 'beside', 'before', 'by', 'near' may be better. 2,3,5
10	The equivalent German construction is close to the English. 1,1,1
11	The German construction is close to the English. 1,2,3
12	The equivalent German construction would not require any article. Test-takers will probably know the English construction 'even as a (noun)' 1,2,2
13	'be allowed to' is an English phrase which all should know. 1,1,1

14 This could refer to 'number' of games, or their variety or the order in which they took place. The German equivalent constructions would be quite close to English.
1,1,3

15 The equivalent German construction is not so similar to English.
2,3,3

16 If test takers know that discus and javelin throwing are field events then 'such as' is the obvious choice. This is an English phrase which all should know.
1,1,1

17 The definite article could also be used in the equivalent German construction, but so could the equivalents of 'some' or 'many'.
'Some' sounds better here if you are translating from German.
1,2,2

18 This may be filled with a type of sport. Some may be unsure about English comma placement.
1,1,3

19 'Varied' may be taken as a verb here, so some may choose 'that'. Not all test-takers may know 'varied' as an adjective. 'Tests of various abilities' would have been easier.
2,3,4

20 The definite article could also be used in the equivalent German construction, but so could the equivalents of 'each' ('*jeder*')
2,2,2

21 Only those familiar with the pentathlon ('*Fuenfkampf*')will know this. Is it jumping?
3,3,5

22 The definite article could also be used in the equivalent German construction.
1,1,2

23 '*Gewidmet*' would be possible in German. Not all test takers will know the English translation of this. 'Devoted' or 'dedicated' would be the best choices.
2,4,5

24 '*den Helden des Tages*' would be the German equivalent. Some may choose 'for' instead of 'of'.
1,2,2

25 Either 'day' could be repeated here, or a pronominal construction could be used. It's hard to put in one word while making the sentence sound natural.
2,3,5

26 Is 'set apart' correct? This may be difficult.
I'm not sure myself, but I think 'set aside' is correct.
3,4,4

27 You just have to add one day to 'fourth' to get 'fifth', or 'next' is also possible.
1,1,1

28 The equivalent German construction is close to the English.
1,1,1

29 Only those familiar with the sports of ancient Greece will know this. Maybe '*Lorbeer*' was used. Not everyone will know the English translation. A more general word may be used, such as 'plants' or 'herbs'.
4,4,5

30 The equivalent German construction is close to the English.
1,1,2

31 'The' or 'that' would make sense here, so the equivalent German construction should be close to the English.
Earlier the passage mentions "very important foot races", so some may choose 'each' or 'every'. It seems impossible for a year to have more than one name, however.
2,2,4

32 The definite article would also be used in the equivalent German construction
1,1,1

Figure 4.7: German L1 consultants' predictions of task dificulty

Japanese L1 consultants' observations about the passage

General comments made by Japanese consultants paralleled those by German consultants, except that the baseline of difficulty, as it were, appeared to be somewhat higher. The average modal difficulty rating awarded by Japanese consultants was 2.5 as opposed to 1.9 for German consultants. If it is acceptable to perform an independent samples t-test on modal values (which are, after all, also a measure of central tendency), then the difference between German and Japanese consultants' difficulty ratings is significant ($t=-2.138$; df62; sig. at <0.05)

Perhaps not surprisingly, of the deletions which receive equal difficulty ratings from both groups of consultants. i.e. (10), (13), (16), (17),(24), (25), (26), (27), (28), (30) and (31), the majority (though not all) fairly clearly on knowledge of 'grammatical' rules. Comments about individual deletions were also elicited from Japanese L1 consultants, and these are shown below:

Del.	Consultant comments (Japanese L1)
1	It should be clear that the missing word is a number, or a word like 'few'. The construction 'every ___ years' is taught in school English. Test-takers may not know that the games took place every four years. We now have summer and winter games, so some may choose 'every two years'. 2,3,5
2	This is difficult. [Only one consultant supplies 'lost'. Following researcher's supply of same:] We learn phrases like 'lose one's way' or 'lose a friend' in school English. The Japanese expression '*naku naru*' has the idea of 'got lost.' 3,5,5
3	Between 'local' and 'international' the [logical] choice is 'national'. 2,2,3
4	'Against' is the obvious choice. The [combination] 'rule against' or 'law against' is taught in school 1,2,2
5	The construction 'No one knows' is taught in school English, but it is usually written as one word with a hyphen. 2,2,3
6	This needs the definite article 'the'. This is taught in school English, but may be difficult as Japanese lacks articles and Japanese speakers find their use difficult. 2,2,4
7	'*Ouyake no shorui*'. 'Documents' or 'records' would fit here. The Greeks did not use paper. 3,4,4
8	'Took place' is an English phrase which all should know. 2,2,3
9	Some preposition of location like 'at', 'near' would fit here. In Japanese you would say '*Orinpasu (mae) no heiya*' ('The plain 'of' or 'by' Olympus'.) 2,3,4
10	'Came' or 'arrived'. Some may choose 'watched' but that is impossible [because there was no television then.] 1,1,2

11 This one might be difficult. Some may not know 'admit' in the meaning of 'allow to enter.' 'But' shows that a negative construction should follow, but some may not notice this. 'A' or 'the' may be chosen because the following word is a noun. Articles are a difficult aspect of English for Japanese. 2,4,5

12 A' or 'the' may be chosen because the following word is a noun. Articles are a difficult aspect of English for Japanese. 1,2,3

13 'be allowed to' is taught in school English. 'Persons' is often learned as the plural of 'person.' 1,1,2

14 'Kinds'?,'range'?,'number'? This is difficult if students [sic] try to choose the best word from those they know. 1,2,4

15 'Include' is learned in school English. Some may choose 'including' because we see it more often in that form. You need to read to the end of the sentence to know that it should be past tense. 3,4,4

16 Most should know that discus and javelin are examples of field events. The phrase 'track and field' is used in Japanese English. We learn the construction 'such as' in school English. 1,1,2

17 'The' or 'some' seems to fit. 2,2,3

18 This may be filled with a type of sport. 'Also' would be close to Japanese '*Nani nani mou arimashita*' ('There was also such-and-such.') 2,2,2

19 Not all test-takers may know 'varied'. This item may be difficult. 3,4,5

20 The definite article is needed here. The construction 'the [such-and-such] of which' is learned in school English. 1,1,2

21 Only those interested in athletics may know this. The pentathlon sports are taught in school. 3,4,4

22 The definite article is needed here. 1,2,2

23 [Only one consultant supplies 'dedicated'. Following researcher's supply of same:] Japan has many national holidays dedicated to certain things [the ocean, sports, the aged etc.] but we don't dedicate days to people. We learn constructions like 'dedicated to his work' or 'a dedicated doctor' in school English, but that meaning is a little different. This item may be difficult. 4,5,5

24 Some may choose 'for' instead of 'of'. 1,2,2

25 'Day' could be repeated here, but students [sic] may not want to repeat a nearby word again. In Japanese you would [repeat the word 'day'. 2,3,4

26 Test-takers might have learned 'put aside' [as in 'Could you put it aside for me?'] but they may not know 'set aside.' 4,4,5

27 You just have to add one day to 'fourth' to get 'fifth', or 'next' is also possible. 1,1,1

28 This is easy, 'were crowned.' / Everyone should get this correct. 'All the' and the plural noun show it must be 'were'. 1,1,1

29 [No consultant supplies 'laurel'. Following researcher's supply of same:] Only those familiar with the customs of ancient Greece will know this. 5,5,5

30 We learn the construction 'So [adjective] was the [noun] that...' in school English. 1,1,1

31 'The' or 'each' would fit here. There was more than one foot race. according to the previous paragraph. 1,2,2

32 Does this mean the year was named after him? That seems strange. If the year is given a name, there can be only one winner of the foot races. So item 30 must be 'the'. 1,2,2

Figure 4.8: JL1 consultants' comments on processing/cue-uptake

A few observations may be in order here. Firstly, although I made no real attempt to assess their level of proficiency in English, consultants of both language backgrounds drew my attention to every one of those deletions (in the full, 41-deletion version of the passage) could be left unfilled without seriously affecting passage meaning. This seems to imply a reasonable level of linguistic sophistication on the part of those who noted these instances.

German L1 consultants made explicit or implicit reference to the parallels between English and German constructions in relation to approximately half of the passage items (cf. items (4); (6); (9); (10): (11), etc.) These references to L1—TL parallels are often to do with similar grammatical; constraints, such has article usage (6), or noun-preposition pairings (4) but also include reference to dissimilarity of L1 and TL constructions as in (15) and usage differences as in (20). Japanese consultants made the same kind of references in relation only to four of five; this is unsurprising in light of the greater linguistic distance between

English and Japanese than between English and German, and in particular of the lack of any article system in Japanese. One of the lexical items mentioned by JL1 consultants was imported from English into Japanese.: 'track and field' is used in the phonetically Japanised form (and romanised as *toraku-and-feerudo*) perhaps as often as the native equivalent *rikujou*) but such imports into German (of which there are far fewer) were not explicitly mentioned. JL1 references to L1 forms in relation to items (2) and, perhaps, (25) may be intended to underline difference or distance rather than similarity; in other words to suggest that recourse to the L1 by task-takers is more likely to lead them astray than to help them fill the blank appropriately.

One implication of this recurrent reference to a role for L1—TL parallels among GL1 consultants is that they may have assumed that informants would make substantial use of, if not translation, then some kind of recourse to the L1 in their approach to the task—even though only a few unambiguous rmentiond of translation as such (see items (17); (23); (29), above) were made by consultants. What is not clear from my notes made during these conversations is whether translation was seen as a 'recovery' process, or as a 'confirmatory' event. Given that L1 parallels are mentioned in relation to only one of the nine cloze items rated higher than 2.5 by GL1 consultants on the 0—5 scale of difficulty, but to 14 out of 23 below the mid-point, it may be that the L1 constraint is thought to offer confirmation or support in most cases. I subsequently put this question to those of the original GL1 informants with whom I was still in regular contact, and received

two replies which were in substantial agreement. In (my translation of) the words of one:

> "It hardly seems possible that [GL1 informants] will not use their native language, even if not consciously. Almost all of them will have been taught about English grammar in German, after all. [...] They will have learned that German and English are [related] and have many similar grammatical constructions and so on. [...] I would expect that for longer or more difficult gaps you will find more translation and [carried out more explicitly] while for less challenging gaps [use of the L1] will be more passive and essentially [be used] to check ideas."

If there is a consensus to be drawn from the figures and comments above, it may be that the task-taker's L1 hovers, as it were, in the background and influences most items in one way or another; in most instances the role of the L1 is secondary. If it is possible to discern a pattern in only five or so references to L1 forms made by Japanese consultants, it would be that three of these pertain to items ranked above the median on the scale of difficulty. JL1 informants tended to rate items as more difficult overall, however, so perhaps not too much can be read into this. More interesting is the lack of any explicit reference by JL1 consultants to translation *per se*. This may seem curious, given that the think aloud and other data reported in chapter 6 offers no real evidence that JL1 cloze task-takers have less recourse items than their GL1 counterparts to the L1. Again, my notes of the conversations with JL1 consultants shed little light on this issue, and so I contacted three to ask about their picture of the role of the L1. Their comments indicated that they would expect JL1 test-takers to translate passage content when faced with difficulty in comprehending it, whether that difficulty was occasioned

by unfamiliar words or phrases, or complex constructions, or the presence of a blank. This, I inferred from consultants' remarks, was almost too obvious to mention. If a consensus can be drawn here, then, it seems to be JL1 consultants see the L1 as playing a more central role in the recovery of fillers, rather than in confirming them. Again, this function for the L1 seems to be in keeping with the greater distance between English and Japanese.

Other consultant observations

Whereas German consultants often implied that test-takers would or 'should' know a particular English construction, or noted parallels between English and German syntax, Japanese consultants made repeated reference to 'school English.' While German pupils study English at least as much as their Japanese peers, for the latter group school English has until recently been by far the most significant source of exposure to the language. In the last few years, cable radio and television and the Internet have begun to offer much greater exposure to English, at least for urban Japanese. The rote memorization ('overlearning'?) traditionally used in Japanese English education means that lexical items taught tend to be remembered in that context, even though they may not be readily accessible in free production. The fact that Japanese assessments of item difficulty are either (34%) equal to, or (66%) higher than, those made by German consultants may reflect the greater linguistic distance between Japanese and English, the arguably lower exposure to the language among Japanese, the greater emphasis on grammar-translation in English education in Japan, or some combination of these

factors. It may, however, also reflect the Japanese cultural trait of what I will label 'preventative humility': "If I tell you in advance that I can't do *x*, I won't lose face by failing."

Comments on scoring and test-wiseness

Consultants from both groups (and this point was raised by well over half of those who volunteered comment) noted that the scoring procedure used—in technical terms the 'exact word' or 'semantically acceptable' options would affect how difficult a given item might be. (My notes of these conversations imply that German L1 consultants' comments on SEMAC scoring were more positive than those of their Japanese L1 counterparts, but this would need to be investigated in much more depth for a firm conclusion to be possible.) Last but not least, consultants' comments seem to offer some insight into the 'test wiseness' informants may, for better or worse, bring to the task. Japanese consultants noted that JL1 informants may be reluctant to fill a blank with a word which has just appeared in the passage, on the grounds that this may seem like 'too easy' an option. I asked a number of other Japanese teachers of English about precisely this point, and they confirmed that Japanese tests of English (typically multiple-choice in format) have often contained 'trick', 'easy' options to trap the unwary. This may well be an example of how prior exposure to certain kinds of test and/or training for tests may affect how quite another kind of task is processed. A perhaps related issue is discussed below, relating to the use of extratextual information.

4.23 Informants' prior expectations about the cloze task

In a conversation with several Japanese researchers in the field of reading and evaluation, I was surprised to hear that one expectation commonly attributed to Japanese students who are to take tests based on foreign-language reading passages is that effectively *all* of the information required to answer the questions will be more or less explicitly presented in the passage. Consultations with teachers at Japanese cram schools (*yobiko* and *juku*) seemed to confirm that this assumption was widely held, and may reflect the strong sense in Japanese testing culture of what may fairly be tested, i.e. information contained within the passage, or taught as part of the well-defined Japanese school curriculum. This realization on my part shed new light on classroom events and previously-constructed semester tests, in which I had noted an apparent reluctance on the part of some students to apply their own knowledge of the topic in drawing inferences or answering questions. Given that the nature of cloze involves the making *un*available of passage information, and the highly unpredictable level of extratextual knowledge successful clozure may require, a cloze task-taker whose approach to the task was conditioned by the expectation that only extant passage content would be of value might well be at some disadvantage.

I could see that it might be necessary to ensure that all JL1 informants clearly understood the potential role of extratextual knowledge in cloze, but I also wanted to find out how widespread this alleged "all-you-need-is-here" expectation actually was. I also wished to find out whether Japanese L1 informants differed in this respect from their German L1 counterparts, and if so by how much. Unable to

obtain data from GL1 respondents by other means, I prepared a short sample English cloze passage (SPILLWATCH) and on an accompanying sheet of paper the instruction below, printed both in English and in German:

"This is a test in which you have to write one English word in each blank space. Please look briefly at the text, but do not write in the missing words. Then mark an X on the line on the pink sheet to show roughly how much of the information necessary to fill in the missing words you would expect to find in the text."

Two deliberately implausible examples followed, showing X marks by someone who thought that almost none of the information required would be found in the passage, and by someone who thought that effectively all of it would be found there. The 'blank' lines (printed on individual coloured sheets) looked like that shown below:

0% ---------- 50% ----------100%

Figure 4.9: Response-line on which expectations were to be marked 'X'

The 50% figure was added deliberately with the intention of discouraging respondents from simply choosing the middle value as a 'safe' option, much as some writers have proposed (cf. Cohen & Manion 1994) that Likert-scales be designed so as to have no exact middle value. One copy of the passage (laminated to prevent completion: actual filling-in of the blanks was undesirable as I sought to access people's preconceptions or expectations rather than their conclusions after having completed the cloze) was distributed by mail, along with five copies of the response sheet and a pre-paid, self-addressed reply envelope, to each of

three of my German L1 former students at the *Volkhochschule* Berlin-Brandenburg and to two more former students at a Berlin university, all of whom had I had taught and all of whom had had previous didactic exposure to cloze-type tasks.

Each of the five 'seed' respondent was invited to fill in one copy of the response sheet, but also to elicit the anonymous responses of up to four others "like yourself." Here language background was specified, but not age, sex or other criteria. Seed respondents were asked to approach further respondents individually so as to obviate 'group consensus', and were not to comment on or even look at response slips, which were to be inserted into the reply envelope via a slit. A seal was provided to prevent loss of contents on return. The same request to gather responses on my behalf was made to five Japanese undergraduate students, again all with previous exposure to didactic cloze. These snowball samples gathered responses from 17 GL1 informants and 23 JL1 informants. Unavoidably, I have no information about these informants beyond their language backgrounds, but my assumption is that the in both German and Japanese groups they were predominantly classmates, college friends, or colleagues of the seed informants. How much prior exposure to cloze or similar tasks these respondents had is also unknown, as is the amount of time respondents spent looking at the passage recorded. (One GL1 seed informant added a note to say that most of 'her' four respondents had, as instructed, looked only briefly at the passage before marking their slips.) Responses were collated by measuring the distance of each 'X' along

the line and converting this position to the closest '5% unit' value. The results were as shown below.

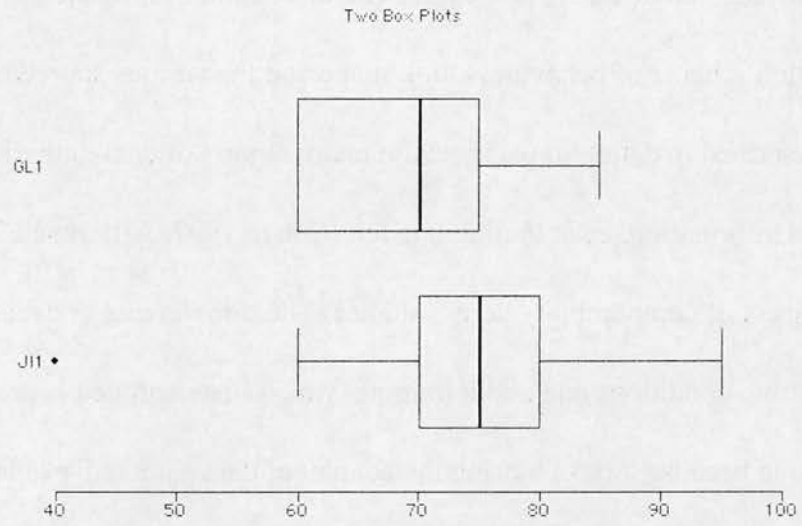

Figure 4.10: Box plot distributions of German and Japanese L1 respondents' expectations of the role of passage content in recovery of cloze deletions

It was assumed that the samples were normally distributed and that sizes (Gravetter & Wallnau 2000:302) permitted the use of the t-test (SPS v2.0.) This indicated a significant difference in the means of the two samples (t-value 2.2763 at df38.) The data gathered in this small survey suggested that, compared to their GL1 counterparts, JL1 cloze task-takers may be slightly more likely to approach the task with a possibly disadvantageous assumption regarding the role of prior knowledge in recovery, namely that this will be less salient to the task. Although the effect on task performance of such an assumption was intuited (that word again) to be quite limited overall in natural cloze, where many items appear to rely largely or totally on activation of cotextual information cues, the finding did lead me to deal more explicitly during informant orientations with the role of extratextual information in cloze recovery.

4.24 Conclusion

In this chapter I have described in some detail issues addressed in setting up a study employing verbal-report procedure. The discussion of how the categorization scheme of behaviours took shape and the various sourced of input into this described in detail–in part because many reports of data-gathering do not convey this information, even though it is felt (Cohen 1997; Afflerbach 2000) to be a key aspect of comparability across studies. The same is true of decisions about reporting conditions and about training, which I have argued is problematic in think aloud because it risks biasing the content of data gathered–or indeed the informants' whole approach to the task. In this regard I am more in sympathy with Ericsson & Simon's original 1984 position, in which informant training in verbal reporting was frowned upon. Finally, I reported a small survey study that suggested that there may be something to the notion that in regard to the role of extratextual information in cloze Japanese L1 informants will have different expectations their German L1 counterparts. As we shall see in chapters 6 and 7, however, evidence for this in the data appears to be slight. In the following chapter I discuss the set of categorizations used in analyzing data gathered via concurrent verbal report, and via its later offshoot, 'annotated cloze', or AC.

CHAPTER 5: VERBAL REPORT PROCEDURES

5.0 Introduction

In this chapter I introduce, illustrate and discuss the derivation of the set of processing events and behaviours used in analyzing the data gathered via think aloud and via two alternatives to that task format. The first of these, so-called 'non-continuous reporting' (NCR) is a verbal report procedure that may best be characterised, depending on the difficulty of the cloze recovery involved each time, as something between immediate retrospection and slightly delayed retrospection (cf. Cohen 1997.) The second, 'annotated cloze' or AC (see chapter 7), replaces verbal report with informant-codings of their own processing behaviours via an 'editable' list of options. I discuss below first a problem that arose in the gathering of think aloud data from Japanese L1 informants (and whose impact was one factor behind the development of NCR and AC) and its implications, I then go on to outline the NCR format, and to briefly compare it with think aloud. This is followed by a discussion of the 'codeset' used in analyzing processing events, and illustrations of its application to data gathered from think aloud informants.

5.1 ..... happens

Many versions were compiled of the 'master list' of processing behaviours observed in think-aloud (see below), and the complexity of some of these speaks of the latitude available to the researcher working largely alone. Some of this complexity stemmed from an earlier stage of data-gathering in which the role of translation into the L1 in informants' cloze processing was of central interest, with the result that a fairly detailed taxonomy of

translation behaviours was called for. Although I will have something to say about translation in cloze in a later chapter, it ceased to be the key line of research I hoped to follow. A brief outline of how this came to be may be of value to those contemplating the use of think-aloud and post-task interviewing as data-elicitation tools.

One of my early conclusions about the cloze processing of Japanese informants was that compared to their German counterparts they made substantially greater use of translation into the L1. Another difference between German and Japanese informants was that the latter appeared very concerned to score highly on the cloze task, whose status was, in the hope of maximizing the authenticity of informants' performance, left somewhat vague. The impression I tried to give was that, yes, this was a test, but one that would have no impact on informants' formal semester grades or informal evaluations. My schedule and those of my informants forced me to conduct a kind of rolling data-gathering, with individual, pair or LL group think think-aloud sessions followed as quickly as possible by post-task interviews (PTIs), and the cycle repeating.

Confidentiality was thus vital, and informants were required to agree not to reveal the topics of the cloze task passages I would or might use, i.e. OLYMPICS and others, or the areas taken up in PTIs. Although the passage topics appeared to remain secure, an informant did (and I believe quite unconsciously) divulge that many of my PTI questions had focused on the role of translation in recovery. Subsequent informants (and because the time of the 'leak' could not be pinpointed

the exact number of those affected is unknown) appear to have misconstrued this focus to mean that translation was somehow an *expected* behaviour in think aloud, or perhaps even a key aspect of what was being assessed. The result was that the amount of translation going on in think alouds from that time on was, to an unknown (but, given the small numbers of informants involved, probably significant) extent artificially inflated. Subtle indications of this in the L1 exchanges between pair-conditions informants were not spotted for some time, and even then not by myself. A good deal of the data already gathered was thus compromised beyond use, for the assumption had to be that any 'additional' translation that was taking place must to some degree have replaced other processing behaviours. If there is a lesson in this unfortunate episode it may be that the period of data-elicitation should be condensed as much as possible in order to reduce the risk of breaches of confidentiality. It might also be useful to have a wide set of equivalent stimulus tasks available; this, however, is not easily achieved with natural cloze passages.

5.2 Implications for elicitation of authentic behaviour

It is tempting to look for some redeeming aspect to all of this, and I propose this: the attempts made by some informants to find out the topics of the data-elicitation tasks (and some also sought this information directly from me) may argue against the notion that authentic data about test-taking behaviours can be gathered only from the products of actual test-taking. The nub of the matter may be less the authenticity of the 'test' task itself, or even its takers' perceptions of it, than of whether or not these individuals' behaviours on that task differ from those they

would employ on an actual test. I will ventilate briefly on this point.

To assume that informants' processing behaviours on data-elicitation tasks such as cloze-centred think alouds are inherently likely to be *in*authentic is to assume that respondents have at their disposal ready-made sets of processing options: "If this were a real test I might read the entire passage before filling in blanks, but as it isn't I'll just get down to it." The notion that respondents amend their processing behaviours according to the perceived authenticity of a 'test' task is not beyond belief, but the weight of evidence in the field of language-learner strategies appears to indicate (Ellis 1994; Cohen 1998) that conservation is more usual: first language processing behaviours spill over into second language processing; the linear habits of early reading disadvantage some readers when they first encounter 'academic' argumentation, etc.

From observations of informants at work on think aloud and annotated cloze tasks (cf. chapters 6 and 7) it was clear that some seemed more relaxed than they might have done in an actual test situation, and of course some, rather than applying separate sets of behaviours according to the task, may have more, or less, seriously applied the processing behaviours they did possess, depending on what they perceived the 'stakes' of the task context to be. (But how can we learn much about this except though some form of verbal report?) Post-task interview remarks by informants gave no indication that (except in terms of such variables as 'seriousness', 'anxiety' and affective response, all of which may vary in takers of authentic tests,) my informants acted differently on data-elicitation cloze tasks

than they would have done on an actual cloze test. Moreover, if the misguided informants discussed above had wished to treat the data-elicitation task as something other than a test, why should they have made attempts to discover the passage topic? The answer may lie in the ubiquity of mock tests in Japanese educational culture, in which a form of 'learning through test-taking' seems to be well established. The potential benefits to studies of test-taking of this zealous mindset among informants will be clear.

Finally, the question of whether data gathered via verbal report tasks can be considered authentic is arguably more vital in the investigation of test-taking processes than in other areas. Reading aloud, translating aloud, and even composing aloud occur spontaneously in some contexts, so that these areas are perhaps more open to think aloud investigation without the question of authenticity arising. There is, however, to the best of my knowledge, no test in use anywhere that requires (or even allows?) its takers to think out loud as they complete it. The corollary of this is that an insistence on *formal* authenticity of the test task pretty much rules out the gathering of data via concurrent verbal reporting..

5.3 A better mousetrap?

Whatever I may have learned from it, the setback outlined earlier in this chapter caused a hiatus in data-gathering, and when I returned to the task it became clear that the set of think aloud behaviour-codings would benefit from revision; several of the codes pertaining to levels and modes of translation (length of translation

unit; accuracy of gloss, etc.) were now less relevant. Equally importantly, I had by this stage come to question the poor return on investment in time and labour that think aloud tasks so often produced. A close re-reading of Ericsson & Simon 1993 appeared to justify the attempt to find less demanding ways of eliciting data about how cloze passages were processed, for these writers' model seemed to have shifted to allow a presumption of reliability even for less concurrent verbal data. This search led me to explore the alternative data-elicitation procedures of 'non-continuous reporting' (NCR) and 'annotated cloze' (AC). Neither of these is entirely original: the former corresponds in some respects with the 'fixed' ('beep' or graphic signal-cued) reporting used by myself in a related, unpublished study, and that described in Cavalcanti 1983; the latter has something in common with the 'checklist' procedures used by Nevo 1989 and Allan 1992, 1995. As I will later compare data gathered via these NCR and AC with that elicited through 'conventional' think-aloud, it is as well to outline them here. The table below sums up the basic features of each:

	Informant task	Reporting condition	'Hard' records	Post-task interview?	Coding by	Coding source	Data attributes
Think aloud	Think aloud continuously about the recovery of deletions and any other aspects of the task	Solo, pair reporting	Audio-recording; task sheet	Whenever possible	researcher	(Earlier) TA data	Often incomplete; often hard to interpret
NCR	Report at each recovery how this was arrived at, and on any other aspects of the task	Solo	Audio-recording; task sheet	As necessary	researcher	TA data	Easier to interpret than TA data; completeness variable
AC	Code recovery operations, etc. via a ready-made list of codes; add written comments as desired	Solo	Task sheet ('manuscript')	Only if necessary	informant	TA data	Easy to interpret; completeness variable

Figure 5.1: Basic features of think aloud, NCR and AC

5.4 Conventional think aloud vs. NCR

The search for a better alternative to conventional think-aloud (think aloud) had two main motives. Firstly, think aloud data is often fragmentary (Cohen 1997; Afflerbach 2000) and for that reason hard to interpret. In terms of the Ericsson & Simon model, of course, this is evidence of immediacy and un-mediatedness, both of which translate into reliability. That said, many of the stimulus tasks used in the development and evaluation of the Ericsson & Simon model appear to have little in common with extended, language-based tasks. Some (cf. the 'Tower of Hanoi' problem) seem to have been shorter, perhaps more open to 'stepwise' processing, and arguably less dependent on the (re)activation of data from long-term memory

and prior knowledge. The easier it is to carry out a task in discrete steps, the easier it may be to report cognitions in a consistently interpretable form, so that in more complex or ill-defined tasks the interpretability of operations may suffer.

Other contrasting features of think aloud and NCR have implications for the interpretability of the data elicited by each. While concurrent think aloud data commonly appears to be self-directed (i.e. the informants has herself as her own main audience; cf. Vygotsky 1986) and NCR directed more at a real or imagined audience, must be careful of overgeneralization. Some NCR informants appear to be addressing a second party (presumably, the researcher) much of the time, while others appear to switch between the two emphases. Think aloud informants, too, may appear at times to be addressing an 'Other', particularly when explaining an inference, or the inability to make an inference or choose between two conflicting notions. The concurrent think aloud informant quoted below was, I think, 'talking to herself' in the first extract, but in the second she is almost certainly addressing an Other, "you know.."; "d'you say…?") while the reference to 'Japanese' suggests that she was addressing the researcher in her own L1:

(1) "[...] some official hmm SOMETHING.. NO [THAT'S] WRONG.. *reports*? (*tsuuchi*).. NO.. DON'T KNOW.. HARD TO DECIDE [...]"
(2) "[...] YOU KNOW THAT THING.. LIKE A.. DO YOU SAY *crown*? IN JAPANESE ITS *GEKKEI-KAN* [...]"
JL1 Sari.

NCR reporting appears to be 'other-directed' to a markedly greater degree than is typical in regular think aloud the case, and this quality tends to make NCR data more readily interpretable to the translator and/or transcriber/researcher.

Side-by-side comparison of think aloud and NCR protocol extracts from the same informants working respectively on OLYMPICS Part (1) and then on OLYMPICS Part (2) will give a flavour of the differences between the protocols the two task formats produce. (To create these tasks, the original passage was simply split into two halves, and the informants instructed to think aloud about her recoveries of the first half, and in the second half to report immediately following each recovery how she had arrived at it.) As above, L1 verbalization is shown in small capitals, and my clarifications in capitals within brackets:

OLYMPICS Part 1/TA: JL1 Takayuki

"[...] WELL.. held every.. WELL.. IT MAY BE EVERY FOUR YEARS.. NO? IN THOSE DAYS IT WAS JUST THE ONE [i.e. ONE OLYMPIC EVENT EVERY 4 YEARS].. SO I GUESS every four years.. [...]"

OLYMPICS Part 2/NCR: JL1 Takayuki

"[....] I PUT jumping.. BUT IT'S JUST A GUESS. IF YOU KNOW ABOUT [THE GAMES'] HISTORY YOU CAN MAYBE GET THIS ONE.. BUT [this filler] IS JUST A GUESS [...]"

OLYMPICS Part 1/TA: JL1 Masami

"[....] WELL.. I THINK IT WAS THE SAME [INTERVAL] .. SO four years [...]"

OLYMPICS Part 2/NCR: JL1 Masami

"[....] I CHOSE jumping BECAUSE I ONCE SAW A DISH THAT SHOWED A MAN DOING THE LONGJUMP [*hashirihabatori*] ... I THINK IT WAS AN OLYMPIC SCENE...

I COULDN'T THINK OF ANY SPORTS APART FROM THAT [...]"

The comparative 'structuredness' of NCR verbalizations will be clear, but the reader will also note that NCR reports may be of shorter duration (with fewer pauses) but need not be comprised of fewer words. Only one trial was made in which two informants were asked to report in the opposite order, i.e. NCR followed by think aloud. In both instances the NCR reporting style largely carried over to the latter task. Both informants claimed to find it difficult to move from

NCR to think aloud, although informants asked to shift in the opposite direction appeared to have had less difficulty. Among informants with experience of both think aloud and NCR, all but two have claimed to find the latter easier. Unsurprisingly perhaps, I have also found NCR data easier to transcribe and, if the LoR was Japanese, to render into English. The same observation has been made by both of the JL1 consultants who have helped me to transcribe and gloss L1 NCR content.

5.5: Does NCR 'lose' data?

But even if the NCR format is in some ways advantageous for both informants and researcher, does it have drawbacks? Given its non-continuity, there would appear to be a risk that pertinent information will, as it were, get lost in the gaps. Side-by-side comparison of (JL1) TA and NCR informants' (i.e. different individuals) reporting of their processing of the same spans of cloze passage suggests that this does sometimes occur (See below.)

OLYMPICS

	TA	NCR	Remarks
Del. 1	"[...] every SOMETHING years [INAUDIBLE] in honour of.. WELL.. every four years I THINK.."	"[...] I PUT four IN THIS ONE.. EVERYBODY KNOWS IT'S EVERY FOUR YEARS"	The TA content tells us that the informant instead inserted a running L1 filler but still filled the blank quickly. The NCR protocol gives no evidence of how difficult recovery was for the informant, although the claim that 'everyone knows' suggests it was very easy;
Del.5–6	"[...] no one RIGHT?.. no one knows how far back *the* [stressed] Olympic games go,, but some [...]"	"WELL.. I KNOW THE EXPRESSION 'no one'.. SO no one knows.. AND HERE [deletion 6] aN ARTICLE IS NEEDED.. so the Olympic games.."	Both TA and NCR protocols seem to indicate that deletions 5 and 6 were processed together. The TA content lacks information about the recovery of either deletion, while the NCR informant explicitly refers to knowledge of an expression in (5) and a syntactic rule in (6)
Del. 11–12	"[...] from all parts of Greece.. but.. a married woman was admitted ..even as a spectator.. [noises].. WRONG.. even as *a* [stressed] spectator.. SOUNDS OKAY.. BUT even as SOMETHING.. IT'S NEGATIVE.. ONLY FOR NEGATIVE THINGS.. I REMEMBER THAT.. SO.. AH but *no* married woman was admitted even as spectator.. NOW ITS OKAY I THINK [...]	2[...] hmhm.. no married woman...hmhmhm.. even as a spectator. WELL.. I LEARNED THIS GRAMMATICAL STRUCTURE [...]"	Here too we find two deletions apparently processed together, with explicit recognition of a TL structure (negative+even) that links the deletions and which both informants recall learning. The content of the TA extract indicates an initial mistaken filler (corrected within a few seconds), while the NCR data appears to be something between TA and NCR. This, along with the flat intonation, suggests that the informant had little or no difficulty in recovering the fillers.
Del. 29	"[...] garlands of wild SOMETHING.. WHAT'S IT CALLED?.. I KNOW WHAT IT LOOKS LIKE BUT [NOT] THE NAME.. WILD LEAVES?.. SOUNDS STRANGE.. IT ISN'T (A) FLOWER(S) THOUGH.. LIKE LEAVES.. wild leafs [sic]? STRANGE.. *GEKKEI* (LAUREL).. WHAT'S THAT IN ENGLISH? [...]"	" [ca.12 second interval since recovery of previous deletion] WELL.. I JUST DREW A PICTURE HERE AT THE SIDE BECAUSE I DON'T KNOW THE NAME IN ENGLISH. IT'S *GEKKEIKAN* (LAUREL CROWN) IN JAPANESE.	The NCR informant's remarks here give only a partial picture of what was going on in the interval of silence that led up to them. We may infer that she attempted to retrieve a TL equivalent for the L1 term but failed to do so. The accurate sketch in the margin of her sheet is further evidence that she correctly understood the filler needed but could not supply it in the TL.

Figure 5.2: Think aloud and NCR data

I take up this question of what gets lost, and how critical this may be for NCR format reporting, in more detail in the following chapter.

5.6 Problems of timing individual recoveries

Where audio-recordings were made of verbal data, it may be possible to calculate the time an informant spent on a given recovery. This is not always as straightforward as one might hope, however. Two deletions may be recovered together as one 'unit', and recovery of a given item may stimulate the apparently immediate filling of one or more others. Moreover, informants interviewed post-task also sometimes claim to have 'mentally' filled a deletion in advance of physically entering the word, even though no unambiguous trace of this may be found in the protocol. As a rule of thumb, a very brief interval between recovery of one filler and that of another may be taken as evidence that recovery of the latter item was not very challenging to that informant, while a longer period of (we must assume) processing indicates greater difficulty in recovery. A comparison of informants' assessments of item difficulty and processing intervals, however, suggests that it may be unsafe to attempt precise associations: a ten second processing time, for example, does not necessarily indicate higher difficulty than one of eight seconds even within a single protocol.

5.7 Some issues in categorising think aloud data

With the exception of some ethnographic studies which entirely reject the idea of isolating and classifying the actions of others, and rely exclusively on 'thick' description, all collectors of verbal report data need to come up with some means of distinguishing and categorizing events or behaviours. The philosophical implications of this are probably lost on few nowadays, having been discussed at length in many texts (Cohen & Manion 1994; Byrne 2002.) What is not always

clear, however (Pressley & Afflerback 1995; Afflerbach 2000) is how the taxonomy or coding scheme employed actually came about, even though this may have important implications (Cohen 1997) for how we evaluate a study's findings. Green 1998:68 notes that "there is little consensus on the precise nature of the coding categories that may be used for the analysis of verbal report data", although, as Pressley & Afflerbach 1995 make clear, researchers' discussion of their criteria for categorization may be so rudimentary that consensus would be hard to achieve. There appears to be a cline of opinion regarding the importance of, as it were, consensus in categorizing processing events. Ericsson & Simon 1993 argue that it is unproductive to try to set up *a priori* coding schemes independent of elicitation tasks, and favour the induction of appropriate categories from the data gathered. Krings 1987:165 makes much the same point:

> "Thinking aloud data cannot be analysed adequately on the basis of pre-established categories. Instead, the analytic categories need to be developed and refined gradually, taking into account the internal structure of the data."

The broad implication of this seems to be that quite different categories may be required for different kinds of data, although it is not clear whether Krings merely wishes to regard think aloud data as different from other data, or whether he would distinguish sub-types of think aloud. Faerch (in Faerch & Kasper 1987) , on the other hand, makes a plea for inter-comparability of coding schemes. This, of course, requires that researchers explain and adequately illustrate their coding categories; as it is, the best we can do (see chapter 3) is may be the rough equation

of researcher A's category *x* with writer B's category *y*.

The pros and cons of different approaches to coding could be discussed at considerable length, but (while acknowledging the difficulties involved) it seems better to state my own approach as simply as possible. Fortunately, perhaps, I had begun to gather verbal report data before I had the opportunity to read widely about the taxonomies proposed by others (A number of these had not in fact been published at that time.) Sceptical of 'armchair' or 'intuitive' schemes, I chose to base my own categories as firmly as possible on data I had gathered myself from German L1 informants. A number of issues in relation to this merit discussion before I introduce the taxonomy of cloze processing operations..

5.8 The unit of analysis

A key element in deciding how to code is choosing *what* to code, i.e. how the flow of verbalization is to be divided up into categorizable chunks. A variety of criteria for the segmentation of language output have been utilised according to the linguistic product under study, and among the more widely applicable of these may be syntactic units (Wolfe-Quintero et al 1998) and the 'communication unit' (C-Unit) proposed (and defined via essentially syntactic criteria) by Robinson 2001. But just as Krings 1987 and others insist that think aloud data cannot be made to fit categories designed for other purposes, so it may be with units of analysis.

Two features of think aloud data–although this does not apply to the same degree to NCR data–create difficulties for C-Unit and other schemes of division: think

aloud data tends to be both incomplete, meaning that absent content must be inferred, and also highly elliptical, so that inferences must be made in order to link content to likely referents, etc. The above is especially true in the case of Japanese-language data, for Japanese is a notoriously 'high context' language (Maynard 1997) in which referents are very often implicit even in normal conversation. Because Japanese utterances are often 'condensed' (ibid.) I identified stretches of JL1 think aloud data in which almost every 'word' could be classed as a C-Unit by Robinson's criteria. The resulting analysis seemed cumbersome:

"[...] AH / CHIGAU / EH? / II KA NA? / AH / THAT'S WRONG / EH? / ISN'T IT OKAY?
['NOTICING' MARKER] / [INFERS ERROR] / [HESITATION MARKER] / [REVISES INFERENCE STATUS > POS.]

The end result of my search for a theoretically well-grounded yet still usable unit of analysis was the conclusion that none of the concepts I had looked at seemed to be much more practical than the 'naïve' segmentation I had been using. This involved simply treating each cloze deletion (with associated context) as the basic 'working unit'—a notion not incompatible with the fact that informants might bring together two or more deletions for whatever reason (cf. the linked recovery on structural grounds of the OLYMPICS span '[...] no married woman [...] even as a spectator'. On a smaller scale, i.e. within the basic deletion unit, discrete processing behaviours might be identified and individually coded. On a more global level, it seemed reasonable to describe any larger 'chunks' of the passage as beginning and ending with their outermost deletions , or the sentences that

contained them. Where recovery of a blank was cued by passage information at some distance, that information could be identified as lying within a given 'chunk'.

The only additional formal units that appeared necessary were 'the passage (as a whole)' and the paragraph–which informants' on-task and post-task comments suggested was a psychologically real entity;

"[...] and here in the next paragraph it talks about races again [...]"

and which was anyway integral to measures of target-cue distance based on Bachman 1985. Extratextual information can also be seen as made up of units of a sort, for although we mighty infer from an informant's protocol a more or less thorough prior knowledge of the passage topic, we can only code those elements of that information which she actually refers to. In pair-condition reporting, of course, the conversational turn was also a relevant unit boundary. In short, this simple, by some standards even atheoretical scheme of data segmentation appeared adequate to the task. No significantly finer-grained but still viable alternative has so far presented itself.

5.9 Inter-rater and intra-rater reliability in coding

Cohen 1997 is only one of many writers to draw attention to the need to assess the reliability with which verbal report, like other data, is categorised. To take intra-rater reliability, or consistency, first, I made a point of setting aside printed copies of informant protocol data both with and without the codings I had applied;

this made it possible to compare my later re-coding of the same data with that conducted earlier. The internal consistency of codings was just over 70%, and the most common discrepancy between codings and re-codings was that processing events uncoded on the first attempt were coded on the second. This may be explicable in terms of greater experience and/or exposure to processing behaviours..

But if I was reasonably satisfied with the consistency of my own coding, comparison of these with codings by others proved more awkward. It must be admitted that my consultation on this point was rather a patchwork affair: only a very few individuals had the time or willingness to acquaint themselves sufficiently with the categories I was using to comment usefully on how I had coded transcribed spans of protocol data. The most common consultation format was one in which I and the other party read over the transcript together and he or she queried any codings that seemed ill-grounded or otherwise open to question. I derived some reassurance from these conversations as well as useful insights, but the imbalance of (supposed) expertise, relative stakes in the outcome, etc. meant that only in isolated cases did I feel there had been a genuine 'resolution through discussion' of the kind that, in many published reports, ostensibly resolves all conflicting interpretations.

I do not think that the ratio of 'agreed' to 'differing' interpretations between my consultants and myself can usefully be quantified, except to say that the former (again, I am not blind to the artificiality of the context) certainly outweighed the

latter. The discrepancies that did occur were too various to list here, but a recurrent one centred around what we might call the 'She did it vs. She must have done it' problem. Put simply, I wished to code processing operations solely on the basis of extant data, so that the filling of a syntactic deletion like '...the winners ..... crowned with holy garlands...' was only coded as 'Gra.' if the informant had made some explicit reference to a grammatical rule, offered an example of copular+main verb, or otherwise 'explained' the recovery in grammatical terms. Some consultants thought this unreasonable, taking the view that there was no plausible way that the blank could have been filled *except* by reference to syntactic knowledge. My counters were that explicit reference to syntax rules *was* sometimes present in protocols, and it seemed counter-productive to mask the presence or absence of additional information. Moreover, while insistence on the principle of coding only extant behaviours might appear trivial in cases like the above, there could well be instances in which making assumptions about non-explicit behaviours would be problematic.

5.10 Another measure of reliability?

Apart from inter- and intra-rater reliability we can find make use of another measure of how accurately or reliably codes have been applied to data. This has to do with the triangulatory' role of information gleaned from post-task interviews with informants. It must be recognised that this situation is almost inevitably unbalanced: the researcher seeks confirmation that she has understood and interpreted the informant's verbal report data accurately, while the informant may simply want get out of the room as quickly as possible. How much weight one

gives to post task retrospections may depend on the model of verbal report one subscribes to. My own position is that, although an informant's retrospections may on occasion be at odds with her protocol data, in general post-task retrospections either (to some extent) confirm the researcher's interpretation, or contribute nothing at all because the informant cannot recall her processing well enough to comment. This is in tune with the fairly well-established finding (Ericsson & Simon 1983; Taft 1991) that once a problem has been solved the route to the solution may quickly be forgotten. In short, the compatibility of an informant's post-task comments with the researcher's interpretation is seen as positive support for the latter. Disconfirmation of an interpretation is rare in PTIs, although the informant's reaction may cause the researcher to question her understanding. Simple absence of informant comment is seen as neutral.

The above-mentioned comparison of concurrent think aloud data and informant retrospections might also be seen as a way of validating the codes and codings of processing behaviour. Validity of codings *per se* is not a topic that receives much explicit discussion in the verbal report literature, however, and validity and reliability may not always be distinguished. My interpretation of the comments of Haastrup 1991, Ericsson & Simon 1993, Krings 1996, and Green 1998 is that, insofar as it exists as a separate enterprise, validation is best seen as an iterative matching process tightly bound to the verbal report data to which it is applied, and one in which the taxonomy of codes may be have to be open to at least minor change throughout the research process.

5.11 Representing the data

A number of decisions had to be made regarding how best to represent verbal report data on the page, and–given the quantities of data that verbal report can produce–a certain economy of presentation appeared desirable. Initial attempts at phonetic transcription proved impractical, and anyway there did not seem to be much real value in this level of detail. I therefore opted to follow Krings' 1987 and attempt phonetic transcription only where this seemed to carry particularly important information. One such instance would be where an informant's pronunciation of a word indicated a lack of familiarity with it, an L1 pronunciation, etc. These features are marked off with slashes around a crude phoneticisation using the regular font: /diskoos/, for example, is meant to represent approximately the pronunciation of 'discus' used by a German L1 informant, while /diskasu/ crudely represents a Japanese L1 informant's production of the same item. Where a full word appeared to be contrastively stressed, it is shown in italics. Conspicuous exclamations (.. ah.. that's wrong!..) and question intonations (..eh?..) are represented in the conventional ways.

Unless otherwise indicated, L1 verbalization is shown in SMALL ITALICS. Content inferred by the transcriber, researcher observations, etc., are contained within brackets. L1 synonyms or explanatory terms used within a span of L1 verbalization are italicised:

"[...] I THINK THAT'S WHAT IT IS.. WE CALL IT *GEKKEI* [...]"

Short pauses are shown by two stops (..) without regard to precise length. Pauses over 5 seconds are shown thus: (6secs); (20secs). Due to the phenomenon of tape

stretch on some longer-duration cassettes, these timings should be seen as approximate.

The decision was made to present all verbal report data in English translation, with these reviewed and corrected in cooperation with native speakers of the informants' native languages. Exceptions were made where an informant's original phrasing was thought to be open to more than one interpretation; in these instances the original phrasing follows in parentheses. Informants' verbal data was edited to leave out material (asides to the researcher, requests to open a window, etc.) with little or no relevance to the recovery task, and this appears in brackets. Verbal data is otherwise unedited except (1) where–mainly in pair-condition reporting–an informant herself requested this, and (2) where a highly distinctive expression or manner of speaking was altered to maintain the informant's anonymity. In a very few instances, overly vulgar expressions used by informants were also excised. Gestures and paralinguistic features are 'described' within brackets: [LAUGHS]; [INDICATES AGREEMENT]; [HEAD MOVEMENT SUGGESTS DOUBT], etc.

5.12 Reading aloud not separately represented

Almost all think aloud informants (and many of their AC counterparts) 'read aloud' spans of the passage, and this behaviour may be assumed wherever several verbatim passage words appear consecutively. For a number of reasons, I have not attempted to isolate or deal in depth here those passage spans which informants read aloud. Firstly, informants' remarks (which jibe with my own self-as-subject

experience of think aloud) suggest that reading aloud is largely an epiphenomenon of the think aloud procedure, corresponding to the sub-vocalisation that may accompany reading in conventional settings, and thus not necessarily directly contingent on the recovery task. Reading aloud during cloze processing, moreover, may occur at a variety of speeds, and in various intonations and span-lengths, which may not be directly comparable. Some think aloud informants were heard to read in a low monotone and at fairly high speed: one male informant explained this behaviour as having to do with the "need to keep going" and "rhythm"– reflections, perhaps, of a notion that physically performing the sounds and intonations of English was somehow beneficial to comprehension and/or production.

Some sort of belief along these lines appeared to be widespread among the students from whom my informants were drawn. When I asked a group of twelve or so JL1 undergraduates (who were, for scheduling reasons, ineligible to become informants) whether they read aloud when working alone on foreign language passages, all who replied said that they at least sometimes did so. Some noted that they also tended to sub-vocalise when reading in class. They knew this, they said, because neighbours would sometimes joke about their behaviour, or ask them to be quiet. Almost the entire group seemed to feel that such 'performance' of a foreign language was useful, with many nodding at the remark of one participant who suggested that it "helps to listen to yourself, at least until you feel more [...] comfortable."

When interruptively prompted (while in the process of reading aloud) with the question: "Who are you talking to now? To the microphone, or to me, or to yourself?", informants typically remarked that this behaviour was rather self-directed than primarily intended for the microphone or for myself as auditor. Some informants read aloud in a fairly natural and smooth manner more or less the whole time, though not always at the same speed. Others veered between 'speed-mumbling' and a slower and sometimes exaggeratedly clear enunciation. (This range of speeds and enunciations cannot, I think, realistically be shown in transcripts.) Prompted post-task (and on a few occasions interruptively) to explain this shift, informants typically agreed with my suggestion that it might be associated with processing difficulty and/or more focused attention. I formed the strong impression while observing informants at work that they tended to articulate more slowly and/or clearly when attempting to recover deletions or to grasp the meaning of extant passage text. As a rule of thumb, I conclude that some think aloud informants appear to verbalise more slowly and/or clearly when experiencing some difficulty in the task. This is a feature informants have seldom spontaneously mentioned while listening to their audio-recordings in post-task interviews, and so it may not be entirely under conscious control. It is also true, however, that informants often appear as uneasy at hearing themselves on tape as many of the rest of us, and may prefer not to draw attention to (what to them may appear as) idiosyncracies of speech.

5.13 Describing and categorizing think aloud data

The construction of the set of descriptors of cloze processing behaviours was, as may already be clear, an iterative process. I drew first on the data I had produced in my own think aloud recovery of German-language cloze passages, and compared the processing operations I identified therein with data from my earliest German L1 informants. This process continued until 'saturation' (Glaser & Strauss 1967), at which point the scheme appeared to describe, albeit perhaps with question marks attached, all the data gathered. (I make no claim to theoretical 'groundedness' (ibid.) in the full sense, however.) It must be kept in mind that the codeset represents the sum of all informant operations identified, so that no one informant is likely to employ them all (Some relate only to pair-condition reporting, and are not available to solo-condition informants) or even very many of them. Some operations are common (translation; reference to knowledge of TL phrases or other 'chunks') while others are rarer (explicit reference to a filler 'looking right'.) Cloze processors differ not only in the kinds of operation they employ in the task, but also in the number of operations they appear to have at their disposal. Perhaps more importantly for our purposes, they may differ quite markedly in the extent to which they are conscious of their processing behaviours. Note here that I do not imply that individual task-takers possess fixed or immutable sets of processing options; although it seems likely that individual 'preferences' or 'first choices' apply, sets of processing operations may vary with the task or other factors that perhaps only a longtitudinal study could adequately record.

5.14 Illustrating the application of the codeset

Pressley & Afflerbach 1995 decry the failure of some other researchers to explain and illustrate the application of their taxonomy to real data, but do not devote much space to illustrating their own scheme. Rather than fall into the same trap , I will try to exemplify via GL1 and JL1 informant (think aloud and NCR) protocol data each of the categories (listed also in Appendix 2) which is open to illustration. Some behaviours, such as reading part or all of the passage ('Rpt'; 'Rall') before beginning to fill blanks, or 'skipping through' the passage, cannot be shown with any economy, for the evidence for these typically takes the form of direct observation, informant retrospection, a longish period of silence and/or barely audible verbal data, or (in the last instance) the protocol *in toto*. For the sake of consistency, I have taken the following examples of codings as far as possible from the protocol of a single solo-condition GL1 informant (Detlef), supplemented as necessary with examples from the protocol of a solo JL1 informant (Yasuko) and from that of a single GL1 pair (Fred & Anneke.)

The appropriateness of individual codings may at times be open to question, which may be unavoidable given that the coding decision was in almost all cases effectively my own. (Recall my comments earlier in this chapter about the balance of power and motive in informal or not-authentically-collaborative inter-coder consultation.) In some instances a degree of confirmation was available from informant's post-task retrospection, but I have avoided explicitly labeling these codings as somehow 'safe'. It must be kept in mind that, while the retrospective information is (much like verbal report as a whole) better than nothing, it still

adds up to an individual trying to describe (in either her, or the interviewer's, second language) what she thinks she can remember or reinterpret from her taped verbalizations about a challenging and fairly drawn-out second-language task.

Codes denoting use of local item context information:

Gra

The informant refers to or applies knowledge of a TL syntactic ('grammar') structure of which the missing items may be part (article--noun sequence, auxiliary verb--main verb, etc.) or which otherwise conditions or explains the choice of filler.

'..just it.. worthwhile.. it REFERS TO DISCIPLINE [i.e. pronoun + referent]' (GL1 Detlef)

Col

The informant refers to or applies knowledge of a TL collocation to which the missing item may belong.

'..richly rewarded.. their state.. government government *YESYE*s.. it goes together' (GL1 Detlef)

Phr

The informant refers to or applies knowledge of a TL phrase of which the missing item may be part.

'on the first and last day [...] yes on the first and last day.. FITS TOGETHER..' (GL1 Detlef)

Idi

The informant refers to or applies knowledge of a TL idiomatic expression of which the missing item may be part.

'sacrificial offerings to the heroes of the day.. WE HAVE THE SAME EXPRESSION [in German].. ' (GL1 Detlef)

SS (SSb, SSf)

The informant refers to information contained within the same sentence

(SSb preceding the deletion; SSf following the deletion.)

'..compared is already there so no gerund [can go] at the beginning ..' (GL1 Detlef)

SP (SPb; SPf)

The informant refers to passage information from within the same paragraph

'then it's not a plus [...] international is coming in the [...] next sentence'
(JL1 Yasuko)

EP

The informant refers to passage information from an earlier paragraph to that containing the deletion.

'although Olympic winners.. yes it talked of winners before [...] winners received no prize money they were in fact..'
(GL1 Detlef)

LP

The informant refers to passage information from a later paragraph to that containing the deletion.

(On a few occasion this was inferred from observation of an informant's behaviour, but no unambiguous reference has been located in TA protocols to information in a paragraph succeeding that containing a deletion.)

KOW

The informant refers to extratextual 'knowledge of the world' which she possesses, or thinks she possesses.

'held every.. *six* years I think..'(GL1 Detlef)

LAN

The informant refers to her knowledge of, or poses a question about, some aspect of the structure of the TL, her L1, or an L3

'..there was boxing wrestling [INAUDIBLE] also also also.. there's too many also's no.. 'cause the one before was.. I think it's also..'
(JL1 Yasuko). [In PTI the informant indicated that she felt, apparently on stylistic grounds, that in English the same word could not appear again here.]

L1Par

The informant refers to a parallel or (near-)equivalent phrase or structure in her L1 (L3Par if the phrase or structure is in a third language)

'no one knows exactly how far because it says but later so no one KNOWS *DAREMO SHIRENAI KEDO...*'
(JL1 Yasuko)

Codes indicating partial success:

WC

The informant cites a word class (noun, verb, etc.) to which the target item must *or cannot* belong.

'well it should.. should be verb here..' (JL1 Yasuko)

L1Eq

The informant inserts an appropriate L1 equivalent for a TL filler.

' some official ehmm some official ehmm *SCHÄTZUNGEN* [estimates] WAS IST [what's] *SCHÄTZUNGEN* [in English]
(GL1 Detlef)
'I think BECAUSE OF [wegen] but I don't know the translation so I put it in German '
(GL1 Detlef)

Und

The informant claims or in some way demonstrates understanding of the meaning of the missing item or its immediate context, but cannot produce a suitable filler.

(Accurate sketch of victor's crown in margin of task sheet JL1 Yasuko)

SWA

The informant indicates that she is seeking a *single-word* filler for the deletion.

'..because of its would be two words [...] due to also [...] I can't think about another word..'
(GL1 Detlef)

Codes denoting 'routes' to recovery or comprehension:

???

The informant either 'passively' gives no indication of how she recovered a filler word, or states that she can give no indication.

'so great so great honour so great was the honour that.'.
(GL1 Detlef)

Log

The informant makes a logical deduction or inference apparently based on passage information.

'..winner of which excelled in running.. there should be another competition in [deletion 21] ' (GL1 Detlef)
'no its not the first and last day because there won't be any victors on the first day..'
(JL1 Yasuko)

LNK

The informants appears to process 2 or more deletions together, or to link them within a single span of passage.

'but no married woman was admitted even as a spectator.. it's no because I know the what d'you say? the structure? no or not and then even..' (JL1 Yasuko)

SKP

The informants reads through the passage at speed, apparently filling only those blanks whose fillers can be recovered quickly or easily. (This rare behaviour cannot be economically illustrated.)

GS

The informant indicates that she is guessing at appropriate filler word.

'just to fill in something because I don't know what to put in I take one of these above [...] but I think I didn't get the sentence so it should be not the best ..'
(GL1 Detlef)

Codes indicating processing in L1:

Tr

The informant translates or glosses passage information within the same sentence.

'..of sacred wood HEILIGES HOLZ [sacred wood (timber)]'
(GL1 Detlef)

Tr+

The informant translates or glosses passage information beyond the sentence.

'slaves women and dishonoured persons were not allowed to compete.. WELL NO MARRIED WOMAN WAS ALLOWED EVEN TO WATCH... GOES WITHOUT SAYING [...] THEY WEREN'T ALLOWED TO TAKE PART' (GL1 Detlef)

Codes denoting evaluation or selection among fillers:

CBA

The informant indicates that she is attempting to choose between two or more candidate fillers.

'..a winner of which excelled in running [...] think a [...] the.. oh I don't know about a and the..' (JL1 Yasuko)

'hard discipline? strict discipline? for the tenth month training..' (GL1 Detlef)

JSR

The informant indicates that a filler 'just seems right', i.e. she does not mention sound or appearance.

'special test with varied ability? [...] it's ok I guess.' (JL1 Yasuko)

Sou

The informant explicitly refers to 'sound' of a filler in evaluating it.

'I think I'll put whose in *the winner of which* sounds very strange..' (GL1 Detlef)

LK

The informant explicitly refers to the 'look' of a filler in evaluating it.

'leaves looks funny but I think it's like that [i.e. correct] .'.
(GL1 Detlef)

Codes indicating task difficulty:

RF

The informant inserts a temporary 'running filler' such as a TL or L1 equivalent of 'something', a sound, etc.

'the third day was something to sacrificial offering to the heroes of the third day..'
'and the fourth day *NANTOKA* of the full moon was..' (JL1 Yasuko)

UPI

The informant indicates unfamiliarity with an extant passage word, phrase, or structure.

'..varied abilities such as /pensalon/ [pentathlon]..'
'..of which /exceeled/ ex'cellent in running..' (GL1 Detlef)

+Diff

The informant indicates considerable difficulty in thinking of a suitable filler, or in comprehending an extant passage span.

'I think I do not quite get this sentence..' (GL1 Detlef)
'[I] don't understand the meaning [...] the year he won the race got this guy's name?..'
'no this one I don't know.. can't think of anything.' (JL1 Yasuko)

Nec?

The informant questions whether a filler is actually necessary.

'richly rewarded by their state [...] government? by their state no? do I need anything here?' (GL1 Detlef)

LFN

The informant announces that she will leave the target item for now.

'no idea so I think I will leave that out also right now; hmm I think I'll skip that first and then go ' (GL1 Detlef)

GUP

The informant announces that she is giving up on the target item. (This event is only rarely made explicit in think aloud, but retrospectively informants may admit to having had no intention to return to a given item.)

CL

The informant indicates that her confidence in her chosen filler is low

'hmm not so very very very ehm ehm *UBERZEUGT* [convinced] about that ..'
'I'm not so satisfied [with it] but I put it in..' (GL1 Detlef)

Codes indicating perception of production or comprehension error:

Cha

The informant indicates that she is changing a previous filler.

' oh no that's not possible two words.. very bad' (GL1 Detlef)

Ah!

The informant indicates that her understanding of (part of) the passage has altered.

'made ah [INAUDIBLE] I GET IT.. made the *hard* training not the first I DON'T LIKE THAT ANY MORE.. ' (GL1 Detlef)

Codes specific to pair-condition reporting:

CP

The informant consults with her partner or seeks her opinion about a filler or meaning.

'so it had no.. just the olympic games, right?' GL1 Fr+An

Did(actic)

The informant appears to be teaching her partner something (she assumes to be) unknown to her

'.. TO MY UNDERSTANDING THERE'S NOTHING MISSING [...] NO REALLY THERE'S NOTHING MISSING..'
GL1 Fr+An

Diff

The informant indicates that she and partner have different answers or interpretations at that point.

'WELL I DON'T THINK IT'S WOOD BECAUSE THERE'S AN ARTICLE IN FRONT..' GL1 Fr+An

Defr

One informant defers to her partner's choice of filler, understanding of the passage, etc.

'SO YOU THINK IT SHOULD GO IN? YES? IT'S ALL THE SAME TO ME...' GL1 Fr+An

Solo

One informant fills a deletion without consulting her partner, or overtly seeking her agreement. (This event was only traceable via post-task interview and/or reference to task sheet/manuscript)

Other Codes:

Rev

The informant indicates that she will review all or part of her processing of the task to date, or return to a given item.

'okay but now I want to move up again to number eighteen if I can' (GL1 Detlef)

TLPar

The informant paraphrases part of the passage in the TL or summarises her understanding of it.

'..so the year of his victory.. I mean the year he won the race got this guy's name?.. doesn't make sense..' (JL1 Yasuko)

(For AC-specific codes and graphic markup conventions see appendix 2)

5.15 Conclusion

In this chapter I have discussed a number of issues that I think important to the use of think aloud as a data-gathering tool, and given some background information necessary if the reader is to be able to evaluate the data presented in the following chapter. In that chapter I present in full or in substantial part protocol data from a few of my more informative solo-condition GL1 and JL1 informants, as well as from paired informants. These lengthier discussions may be seen in part as further exemplification of coding decisions, and in part as brief 'case studies' of individuals' (and pairs') processing decisions and steps. This section is followed by the presentation of condensed data in tabular form, summarizing the total product of my applications of think aloud procedure. This format allows the tentative identification of any patterns within the data (though it must be said that it is rather underwhelming in this respect) but will also point up some of the limitations of think aloud as used in these contexts. Some points of theory, as well as brief reports of small 'sideline' studies conducted, are introduced at what I hope will seem appropriate points.

CHAPTER 6: DATA GATHERED VIA THINK ALOUD AND NCR PROCEDURES

6.0 Introduction

In this chapter I present in full or in large part the think aloud protocol data of several GL1 and Jl1 informants who reported individually or in pair-condition. My experience of looking at verbal report data gathered by other researchers has been (and cf. Pressley & Afflerbach 1995) that access to longer stretches of protocol data is necessary in order to form a picture of the categories that may later be condensed into tables of frequencies. Hence, and in an attempt to show something of the range of 'informant-productivity' on think aloud, I include the full protocols of three solo-condition informants. These are (GL1) Claudia, and (JL1) Yasuko and Harumi. The last of these, although sparse in terms of insight generated into the informant's processing, does not in fact demonstrate the weak end of the think aloud data spectrum; as mentioned in chapter 4, there seems to exist a sub-population of individuals who genuinely appear unable to report on their cognitive processing to any useful degree. Having almost certainly never been asked to think aloud about a current task before, however, these 'low-verbalisers' may be unaware of their unsuitability to verbal report tasks and may thus volunteer as informants. There is of course little point in reproducing a low-verbaliser protocol here, but were I to do so it would be very brief indeed. I go on to cite extracts from the protocols of GL1 and JL1 pair-condition informants, and discuss after each protocol or extract some of the processing events and behaviours it contains.

In this chapter I also present a variant reporting procedure labelled NCR, and illustrated by the full protocol of one JL1 informant (The task format was not used with GL1 informants.), Ryou (Extracts from the protocols of other JL1 NCR informants are shown in Appendix 4.) I discuss the potential benefits and drawbacks of NCR vis-à-vis think aloud and look at their relative efficiencies as data-elicitation tools. Using tabulated frequencies with which individual deletions were filled, and how these recoveries were coded, I attempt to draw comparisons between the processing behaviours of GL1 and JL1 informants inferred from think aloud data (figure 6.2 in this chapter) and between JL1 informants in think aloud and NCR conditions (For reasons of space, cf. figure 7.14 in chapter 7.)

6.1 Protocol of GL1 Solo-condition think aloud informant Claudia

In the protocol data that follows the basic unit of division (see chapter 5) is taken to be the individual cloze blank, so that the initial segmentation is by deletion number. The points of division between individual deletions in the protocol data are to some extent arbitrary, but are intended to occur at 'natural' break-points or shifts in focus within the informant's think aloud. To make clear which codings have been applied to which spans of data, each coded span is presented on a new line. Insights gleaned from post-task interviews re presented along the data to which they most closely pertain. Comments in past tense ('Yasuko claimed..') were elicited post-task; observations from my notes during the task or from the data itself are in present tense and in italics. Once again, event codings made possible only by retrospective data are underlined, and 'probable' codings are followed by question marks. Rather than use code numbers to identify these 'case

study' informants I use first names, but to conform with the promises of anonymity and confidentiality made to informants, their names have been changed to others appropriate to the informant's ethnic background and sex.

A few more words about informant Claudia may be in order. Recruited into the informant pool via an existing volunteer, Claudia appeared well-suited to the think aloud. Thoughtful but confident, with slightly above average (compared to the GL1 informant pool as a whole) English language ability, Claudia had no hesitation about reporting in solo condition, and stated a preference for using English as her main language-of-reporting. Because Claudia arrived late for the reporting session, I opted to curtail her task to the 32-deletion version of the OLYMPICS passage, although in the event there would probably have been enough time for her to have completed the full 41-deletion version. The application of codes to Claudia's protocol data is illustrated below, along with relevant observations as well as insights gleaned during her post task interview. (NB italicised information in the 'Comments' column relates to Claudia's think aloud performance; everything else refers to her post-task interview.)

NB Except where otherwise indicated at the head of a protocol (extract), L1 verbalisation is shown in small capitals and TL verbalisation ion regular type.

Del.	Segmented protocol data	Codings	Comments
1	the Olympian athletic festival.. held every hmhmmhmm years.. suggests that it should be a number.. I'll say *four* years because it's held every four years today so *four* years	RF Phr/Coll KOW	

2	in honour of Zeus eventually... CAN'T SEE THE CONNECTION.. can't see the connection umm.. maybe it could take a verb related to Zeus but I'm not sure... umm.. I'LL CARRY ON	WC LFN	In her post-task interview, Claudia remarked that the sentence was "very hard to understand."
3	next umm.. became first an hmmhmhmm event.. must be an adjective and..connected to national because later in the sentence there's international.. *national* I'D SAY	RF WC Col SSf	In her post-task interview, Claudia. confirmed the collocational dimension, noting that "[national and international] belong together".
4	and then after the rules.. *against* foreign competitors had been waived international	??? Col	In her post-task interview, Claudia noted that 'rules' and 'against' "belong together"
5	umm then there's the next sentence.. one knows exactly how far back... umm there you have to... I would say *no-one* knows exactly how far back but because *no-one* is written as one word it must be something else... I'M NOT AT ALL CLEAR ABOUT WHAT...	LAN +Diff	In her post-task interview, Claudia suggested she may have been confused by the L1 equivalent (niemand), which she claims "cannot be divided."
6	umm how far back... *the* I would say *the* Olympic Games go	??? Gra	In her post-task interview, Claudia noted that "The noun needs an article."
7	but some official there I would say you need something like *record*s or *picture*s or WELL.. something because official suggests a noun and the noun has to do with the possibility of restricting [the reference] so I'll say some official hmm.. what can it be in English?.. which dates nothing comes to mind WELL I'D SAY [...] OFFICIAL *DOCUMENTS* OR SOMETHING WRITTEN... AND THAT'S WHAT IT RELATES TO	WC UND +Diff L1Eq	*Claudia appears to seek assistance from researcher, and is invited to produce in her L1.*
8	umm.. the games.. place in August on the plain umm.. I'd say it could be something like *always* or yes *always* in fact [...] because it's a kind of recurrence so it must be a time-related word... A SENTENCE COMPLETING WORD... *always* wouldn't be too bad	Log	*Claudia pronounces 'place' as /pleiz/.* She could not recall in her post-task interview if she had interpreted the word as the TL item 'plays', but was aware that her filler was unsuitable. In the PTI Claudia also seemed to have realised that the missing item was part of a phrasal verb, and suggested the L1 equivalent 'statt finden' (take place)
9	on the plain *of* Mt. Olympus?.. because plain is related to.. the mount and so it can go with *of*	KOW? Gra?	Asked in her post-task interview whether extratextual or grammatical knowledge played a greater role here, Claudia answered "Both".

10	and it's *come* I mean *came* from all parts of Greece because there's no verb there	Gra	In her post-task interview, Claudia. claimed to have considered 'gathered', but to have chosen 'came' as it was "simpler."
11	but... okay contrast.. so *no* married woman was admitted	Gra?	In post-task interview Claudia. claimed not to have had prior knowledge of this restriction on women spectators, and to have relied solely on the contrast marker as a clue.
12	even as hmhmmhmm *a* spectator because it's not plural *a* spectator... and nothing else fits apart from the indefinite article	RF Gra	
13	umm slaves women and dishonoured persons *were* not allowed to compete because it says to be allowed and so the auxiliary verb is missing...	Gra	
14	the exact... okay that could be something like *order* or umm *order* would fit... I'll write it in now	??? Tr	In the post-task interview Claudia claimed to have thought first of L1 *Reihenfolge* and selected *order* as a TL equivalent.
15	but the umm of events is... but events is uncertain here.. I'd say you need something like.. umm boys' gymnastics horse-racing field events... umm a kind of similar word... but events *as* well GRAMMATICALLY THAT'S NOT SO GOOD... but events *as* boys' gymnastics	Gra CBA	In her post-task interview Claudia asked what the missing word was. She seemed very surprised to hear that it had been *included*, even though a direct L1 exists; she could not recall why she felt another event could fill the blank
16	horse-racing field events well field events [INAUDIBLE] as /diskus/ and javelin throwing and hmm difficult [INAUDIBLE] field events *as* discus that must... *those* as discus and javelin throwing?... it must have to do with [examples of] field events and then a comma...	KOW LAN	*Claudia appears frustrated by deletion 16.*
17	perhaps *some*... very important foot races because there are various distances so there will be several... *some* or *a lot of* are more descriptive.. a counter would be better.. some... *a lot of*... there was... that's the problem	Log KOW?	*Claudia appears to forget the cloze stipulation of single-word fillers.* (In her post-task interview however she claimed to have seriously considered only *some*, and not *those* as a filler in 17) She added that she had indeed noted the comma, but could not recall whether it had affected her understanding of the sentence.
18	HOW DO YOU LUMP boxing AND wrestling TOGETHER? I'll move on	LFN	*Claudia appears to seek a superordinate term for the two events.*

19	and special tests umm there I'd put *which* because it relates to this varied ability... umm special tests.. umm abilities a noun couldn't go in either...such as pentathlon which varied.. that's no good doesn't fit... or maybe? or? boxing and wrestling and special tests or.. and and.. no it's hard to work out what varied ability.. what it's related to is varied a verb in fact.. is it? abilities is clear... varied abilities would it be?.. if that's what it's related to something.. completely different has to go in there.. maybe I'll get it yet... I'll do the next one	Gra LAN LFN	In post-task interview C. confirmed that she had initially interpreted *varied* as a verb, so that she perceived *special tests* as the subject of the clause. She claimed to have learned the adjectival form as *various*, but confirmed that she had begun to suspect that *varied abilities* might be adjective+ noun..
20	[6secs] *the* winner of which excelled in running	??? Gra	In her post-task interview, Claudia noted "The article was missing." *Claudia uses 'sentence' here in the sense of 'clause'.* (The L1 equivalent 'Satz' is used in both meanings.) Claudia confirmed in her post-task interview that she had been trying to find or evaluate fillers by how correct they sounded, and could see no other grounds for preferring the definite or indefinite article. Claudia claimed not to have known that the pentathlon (L1 *Fuenfkampf*) was the name of an event. Interestingly, she appears to arrive at a correct conclusion, i.e. that the filler must be the name of an event, via this misinterpretation.
21	It needs another noun... discus and javelin throwing and wrestling [...] doesn't fit because it's related to the bit before discus and javelin throwing and wrestling... umm the sentences go because.. together [and] in the second sentence the verb is missing... so it must either relate to the first one.. which excelled in running winner of *the* winner of which of which ... *a* winner which excelled in running such as the the penta.. pentathlon? oh right! the pentathlon is what the last sentence is about.. the winner of which excelled in running... now I've got it! the pentathlon is the athlete.. I mean the winner of the race so maybe [it needs] here another noun... running discus javelin throwing wrestling so there must be another type of sports event here... the winner of all these events is the pentathlon right... so which sports events to put in? maybe wrestling rowing javelin /diskus/ boxing	WC SSb Gra Sou UPl Ah! Log	
22	so evening of the day was to... sacrificial offerings to the heroes... uh.. one evening of the third no *the* evening of the third day the evening is more precise... so it's *the* evening of the third day	Cha ??? Gra	In her post-task interview, Claudia pointed out that "of course" a day can only have one evening, and that by "more precise" she had meant grammatically more accurate.

23	was umm *used*? to sacrificial... sacrificial THERE HAS TO BE A VERB THERE... sacrificial offerings to the heroes	WC	
24	mm *of* the day ... because each day they had one sports event [so] on each day they must have had a winner of the day...	KOW Log	
25	and the fourth day... uh the full moon was set mm as a holy day and the fourth day [INAUDIBLE] the full moon... of the full moon is that THE FULL MOON?.. perhaps its a more precise description of the day.. the name of the day perhaps... *day* of the full moon... a closer description of it?	UPl Tr?	Claudia confirmed that the indecipherable whispering on her protocol represented her attempts at glossing the full sentence in her L1.
26	was set *used* this day this holy day? well umm not day used to use a day? no you don't say that was set as a holy say.. UMM.. ACTUALLY I'D SAY IN THIS SENTENCE THERE'S NOTHING MISSING so it's hard to imagine what could go in here... was set... well in fact it could need another verb or something that more closely describes it	Nec? Diff WC?	In her post-task interview C. again asked for confirmation that a word was missing. Told that the missing item was *aside*, she claimed to know 'put aside', but not to have heard of 'set aside'.
27	on the...yes there has to be a number here and it's the fourth day... the holy day so I assume if the order carries on it 's the *fifth* and last day	Log/SPb	
28	now we need an auxiliary verb *were* crowned	Gra	
29	with holy garlands of wild umm okay I can imagine what this is... probably this crown of what d'you call the plant...it's like a ring [GESTURES] wild... well I'd say wreath...holy garlands of wild from a sacred wood must be something like that.. but wild bothers me a little... crowned with holy garlands of wild... you don't say MEMORIAL WREATH in this sense... anyway it must be something to do with the plant	KOW CL	*Via gestures C. clearly indicates that she knows what the missing word represents and seeks confirmation from the researcher. She creates an L1 paraphrase but rejects it.* In her post-task interview C. made an accurate sketch and asked what this (laurel wreath/olive crown) was called in English.
30	here it must be so great *was* the honour	SSb (see below) Gra?	In her post-task interview C. noted that she had learned the inverted form 'So (ADJ) was (NP) that...' in school.
31	that the winner.. yes it needs a verb.. so great was the honour.. that the winner of *the* foot race because it says the winner *of* and so you need an article	Gra	In her post-task interview C. confirmed that she had momentarily returned to the deletion (30) to confirm the filler

32	gave his name to... *the* year of his victory I'd say.. it would be logical to have *the* year of his victory and this *of* relates to the article	Log Gra	In her post-task interview, C. claimed to have briefly considered 'each' as a filler, but to have decided it was "logical" that a winner would win only once.

Discussion of Claudia's protocol data (above)

Claudia may be seen as a 'productive' informant, for in only six instances did she fail to provide enough information about her processing of the blank to allow coding, and in two of those she retrospected in sufficient detail to provide some insight. Claudia's verbal reporting is also more than averagely structured and thus fairly easy to follow. Close reading will show that her verbalisations are seldom fragmented, and are often in the form of more or less complete sentences. In terms of the classical model of verbal reporting discussed earlier this might be seen as casting doubt on the validity-cum reliability of her reports, but there is no reason to think that the data represents anything other than Claudia's natural manner and speed of speaking, even in her L2. In short, the range of natural and authentic modes of 'thinking aloud' may be rather wider than the 'classical' Ericsson & Simon model seems to allow.

Claudia's palette of processing operations is fairly wide, with a total of 19 identified in her concurrent or retrospective data. She refers to knowledge of grammatical rules or structures in 12–13 recoveries, but to phrasal or collocational knowledge in only three. 'Logical' inferences appear in six items, and extratextual 'knowledge of the world' in four or five. Translation seems to have played a role in only two recoveries, and even this was noted only via retrospective remarks. Although Claudia did on quite a number of occasions give the appearance of

having difficulty with recoveries, explicit reports of difficulty occur in only three instances. In two of these, and in three to four other cases, however, she resorts to mentioning the word class of the missing item: of deletion (21), for example, she notes that "it needs another noun." This device, informants have agreed in conversation, indicates at least temporary inability to settle on an appropriate filler, and their comments strongly suggest that 'narrowing down' to the word-class may simultaneously help the informant focus her subsequent processing and demonstrate (for the benefit of the auditor/researcher, and/or herself) that she has gone at least some way towards recovery. In only three cases does Claudia abandon recovery of the deletion, and she does not return to any of these items.

As for the 'level' at which Claudia's processing takes place, it appears to be largely local. She relies, as already mentioned, quite heavily on grammatical knowledge, and only a few blanks such as (3), (21), (22), and (30) does she appear to mention information from within the sentence but outside the immediate syntactic/phrasal context. Claudia does on occasion apply information from neighbouring sentences, however, as in (27) where she appears to use the mention of 'the fourth day' immediately prior to deletion (26) to logically infer that the next day to be mentioned will be the fifth. This deletion thus carries the 'double-coding' Log/SPb to denote an inference explicitly based on information earlier in the paragraph. There is debate (cf. Oller & Jonz 1994) about how the 'extratextuality' of information relates to the 'level' of processing within cloze and other tasks, but it does not seem plausible that any of the schemes based on

that in Bachman 1985 conceive of this relationship in terms of literal distance beyond the paragraph or passage. However it was brought into play, knowledge of the world appears to have played a role in Claudia's processing of five to six items. This may be a matter or interpretation: in her segment pertaining to deletion (17), below, is she basing her inference on prior knowledge that there were in fact races over several distances, or is she using 'distances' to mean something like 'events', so that no extratextual knowledge is necessarily implied?

"perhaps *some*... very important foot races because there are various distances so there will be several... *some* or *a lot of* are.. more descriptive"

This question was not clarified in the post-task interview (which followed the think aloud task, with only a short break between) and is probably not of central importance to cloze recovery given that a number of informants chose 'some' simply on the syntactic ground of the plural 'races'. The point does underline, however, that coding of processing behaviours is sometimes a matter of 'best guessing'.

In the majority of deletions Claudia appears to have taken up only one or two cues, and it is unlikely that all of the codings applied to, say, deletions (19) or (21) were equally important. (The question of the relative salience of processing operations is taken up gain in the following chapter.) Nor does Claudia explicitly generate multiple candidate fillers in more than very few instances. A remark she made in conversation a few days later with myself and another think aloud informant from the same class, to the effect that thinking of many alternative fillers is "not very

sensible" as it means one then has to choose among them, led me to wonder if this had been a (semi-)conscious strategy on her part. (She had not mentioned this in her post-task interview.) The rest of Claudia's remarks during the abovementioned conversation were in a similar vein: if one could not recover a blank fairly quickly, it was best to move on. This, however, is not something Claudia herself practiced very often–indeed her processing of items such as (19) and (21), above, suggests rather a certain doggedness. The unreliability of 'self-reports' (see Cohen 1998:39ff) distanced from concrete tasks is fairly well-established, however, and I noted the comments above only as evidence of the potential mismatch between 'theory' and practice.

While Claudia's post task interview comments made possible the interpretation of her processing of just three deletions, (4), (6) and (14), which could not be coded from think aloud data, they offered some useful insights into her extratextual knowledge and other aspects of the task and confirmed some researcher inferences. As I have already mentioned too often, the classical model of verbal report sees post-task recollections as inherently less reliable than on-line verbalisation. My impression of Claudia, however, was that she was willing to volunteer in her interview only information of whose accuracy she was fairly confident. On at least eight occasions (according to my notes at the time) she either spontaneously stopped herself from giving more information in response to a prompt, or withdrew a comment she had made. In relation to deletion (21), for example, she seemed to imply that her identification of the pentathlon with the winner may

have resulted from a confusion of the words 'pentathlon' and 'pentathlete'; she paused for a moment and then said: "Leave that out. I'm not sure of it." Claudia's apparent sense of caution inspired confidence in her retrospective comments and elaborations of her earlier processing. My reading in the field of verbal report data has yet to turn up a serious discussion of criteria for weighing the reliability of 'supplementary' retrospective data, and it may simply be up to the researcher to resolve this in discussion with individual informants. Other researchers with whom I have discussed the point have echoed my impression that the retrospections of some informants are *intuitively* more reliable than those of others, but no objective assessment of this seems possible.

To conclude the discussion of Claudia's task processing, although it is tempting to award her some measure of credit for 'nearly' recovering the correct filler (as in item (5), where she then rejects it due to a mistaken notion of TL orthography) , her cloze score according to the criterion of SEMAC fillers actually set down on paper was 66%. Claudia's overall time-on-task (think aloud) was just over 20 minutes.

6.2 Protocol of solo-condition JL1 think aloud informant Yasuko

Yasuko volunteered as an informant on the strict condition that she would *not* be asked to report in pair-condition. Like Claudia, she expressed a strong preference for reporting in English. Although her scores on class paper tests (including didactic cloze tasks) and reading comprehension exercises were average or better, Yasuko's oral production sometimes appeared to deteriorate in challenging

situations such as (semi-)formal group discussions of complex environmental or social issues. For that reason she was reminded that she was free to report in either the TL or in her L1, and to change LoRs at any time. As plenty of time was available for Yasuko's reporting session, she was set the 41-deletion OLYMPICS cloze passage; only data for the first 32 deletions are shown here

Del.	Segmented protocol data	Codings	Comments
1	every four years I know it because the current Olympic is held in every four years so I think this should be true from the beginning but I'm not sure	KOW	Yasuko noted in her post-task interview that this was "common knowledge."
2-4	Eventually SOMETHING its local character eventually something its local character [INAUDIBLE] became first a SOMETHING event and after rules against LEARNED THAT [collocation] IN SCHOOL foreign competitor..had been waived.., so rules against foreign competitor been waived.. international.. so I guess it eventually lost its local character.. I guess became first a national event because I thought first it might it might be international event but then it's not *a*.. plus the international is coming in the next paragraph so.. next sentence.. so [I] think it's the first it lost local character and then it became national then it became international..	RF (L1) RF (TL) Col KS LAN SSf Log?	(*SOMETHING = NANTOKA*) *Yasuko seems to confuse paragraph, sentence and clause.*
5	UMMM.. one knows exactly how far back olympic games go but some official date from.. so no one knows exactly how far because it says *but* later.. so no one knows NO ONE KNOWS far back how far back olympic games..	Gra/SSf L1Par	
6	the? these?.. umm.. I don't know I'm not sure about those things.. I think it.. here comes those [INAUDIBLE] particles.. not particles I mean the and a and stuff.. it's not a so maybe the.. (9secs)	+Diff Gra	Yasuko noted in her post-task interview that she has "always" had difficulty with English articles.

7	but some official date.. official (6secs) some official official something date from 776 BC some official documents (7secs) I guess because.. it says official.. official thing is some information which exist or which is like.. oh legal things so I think it's official documents (5secs) some paper official paper.. no I think documents is better words..	KOW JSR Col	Yasuko suggested in her post-task interview that 'official' and 'documents' "go together".
8	the games SOMETHING place in August.. take place ..take place in August.. take place means it happens there ..I know this take place the word I mean.. what d'you call the words.. IDIOM in Japanese.. because I learned in school like we have list of those things like take part in take place um.. like all other things like two or three words put together and then have meanings its in Japanese called IDIOM.. IDIOM? I think idioms anyway.. when I was school we have to memorise those things so I know take place means the games happens.. happened.. oh it should be past tense yeah? took 'cause it's in.. in early days so it should be all past tense? [INAUDIBLE] yeah.. should be past..ehh.. took place in August	Phr/Col l/Idi KS Ah! Cha SPb	Yasuko noted in her post-task interview that the Japanese equivalent of 'idiom' is applied to English phrases, collocations, etc. ('Expression' might be a better translation for the Japanese term.)
9	on the plain near? mount Olympus.. Olym.. Olympus plain.. oh it should be umm.. preposition here.. DON'T KNOW.. near? at the bottom of? no it's just one word so umm.. plain of.. no because Olympus is a mountain plain is a flat land so it should be near.. or there could be something else but then it's just one so it can't be in front of or.. or other things	WC KOW	Yasuko confirmed in her post-task interview that she knew a preposition was needed, but could not say why she chose 'of'.
10	(9) uh many thousand spectators... came from all parts of .. umm.. I it should they should be verb here FROM GREECE.. well.. it should be come because it says *from* all parts of Greece..	WC Phr/Col	Yasuko noted that 'come from' "goes together".
11	umm.. but married women was admitted as.. even as a spectator.. but no married woman was admitted even as a spectator.. it's no because I know the fact that the.. the.. actually I think no woman was even married.. unmarried was.. was was admit..not admitted.. was umm.. anyway I know that no woman was allowed in there so there should be no here.. plus there's like even ..the word even.. SO [the filler] SHOULD BE no.. NO SOMETHING EVEN SOMETHING in Japanese..	KOW Tr? L1Par	Yasuko noted in her post-task interview that she had glossed this structure in her L1.
12	even as a spectator	???	Yasuko noted "I just knew it."

13	slaves women and dishonoured persons were not allowed to compete.. ‘ cause there should be a verb here and allowed is here and it should.. it’s ..it’s.. it’s umm.. passive so I need here what d’you call.. in Japanese COPULAR like be.. and its past and its plural so it should be that.. should be were	WC Gra/Col L1Par Gra	Yasuko noted that she recalled these labels from school English class.
14- 15	um the exact events is uncertain but events da da da... dates no this is not date.. is that [INAUDIBLE] of events ? umm exact course of events.. no it cannot be course of events ‘cause there are other umm.. kinds of events following this sentence so there should be number the exact number of events is uncertain but events include included boys’ gymnastics horse racing field events such as discus and javelin throwing and uh very important foot races (5secs) also but events included boys’ gymnastics because I think the events need need verb here and then umm.. they say they’re talking about events and then they’re uh.. different kinds of events following so it should be the verb should be include	RF SSf WC SSf Log	
16	and then field events such as.. I know this this this this what do you call this idiom such as da da da [expression introducing] A KIND OF event RIGHT? discus ALSO javelin throwing KIND OF events.. so there should be such as..	Phr L1Par	
17	and very.. important foot races foot races.. what does it mean foot races this like? marathon and stuff? I don’t know why they.. what is this? and *also*? [...] I’m not really sure about this one I think its also.. I cannot think of anything else.. I’m not really sure about what foot races are..	UPI +Diff CL	

18 (14-17)	and uh line fourteen there was.. boxing and wrestling and special test varied abili.. and special test [INAUDIBLE] such as /pantathlon/.. /pantathlon/ (10secs) there was boxing wrestling [INAUDIBLE] also also also.. no there's too many also's,, no 'cause the one before was I think it's also umm.. JUST A MOMENT.. should go back.. the exact number of events is uncertain but events included boys' gymnastics horse racing field events such as discus and javelin throwing and also very important foot races	UPI LAN Rev	Yasuko confirmed that she had felt "uncomfortable" about using the same word again.
19	there was [INAUDIBLE] comes also.. also boxing and wrestling and special test FILL with varied ability? I think it should be preposition here but I'm not really sure /pantathlon/.. /pantathlon/.. I'm not really sure what this is either but.. but I guess it's varied ability it means some.. some sort of game with like.. combination of like running and throwing and jumping and stuff so.. so I think special test with varied ability? [INAUDIBLE] its ok I guess..	WC Log JSR	Yasuko confirmed that 'pentathlon' had been unfamiliar to her, and that she had made a logical inference to derive a meaning.
20	umm.. /pantathlon/ winner of which the.. a winner.. a winner of which excelled in running (5secs) a [I] think a.. .. oh I don't know about a and the.. should be a 'cause its not specific person..	Gra	Yasuko noted that she had applied the guideline that indefinite reference calls for the indefinite article. (cf. Ryou's NCR data in a later section)
21	a winner of the.. ah.. which means /pantathlon/ so /pantathlon/ is excelled in running discus and javelin throwing wrestling.. oh I really don't know what this is.. uh (5secs) so discus and throwing things wrestle run swim?.. no I don't think swimming because I don't think they had swimming pool those days umm.. what could it be? running wrestling throwing oh there may be something umm.. afterwards and cha cha cha no no no there's nothing (7secs) there's /bosksing/ and wrestling.. maybe boxing I don't know umm.. don't know what this is.. maybe jumping high jump? jumping? Jumping.. maybe swimming.. no not swimming.. umm ..ok anyway umm.. the winner of which excelled in running jumping discus and javelin throwing and wrestling	CBA KOW SPf/LP +Diff	In her PTI Yasuko claimed to have searched the remainder of the paragraph and "maybe" beyond it for information about pentathlon events.

22-24	evening of the.. third day.. was to sacrificial offerings to heroes (8secs) in the evening.. no there should be just one *the* evening of the third day was.. because this is,, you don't need in because umm.. evening.. evening of the third day is the subject.. so just the evening the evening of the.. every evening? of the every third day.. evening of the third.. no not evening.. of the third day the evening of the third day evening of every third day.. no the evening of the third day was umm (5secs) to sacrificial offering to the heroes of the day.. but I don't know .. does it.. was.. on the fourth day.. evening of the third day was.. SOMETHING to sacrificial offering to the heroes of the third day (8secs) evening of the third day so [INAUDIBLE] the heroes of the day.. evening of third day was the time when they give this offering to the heroes of the day.. but umm.. so umm.. there should be verb here [INAUDIBLE] evening of the third day was kept? no reserved was uh scheduled.. umm [INAUDIBLE] don't now umm.. not really sure what comes here.. evening of the third day was umm (5secs) was the day of the umm.. no was the [INAUDIBLE] the third day was umm (9secs) set? I do it later..	Gra RFL1 TLPara WC CBA +Diff LFN	Yasuko confirmed in her post-task interview that she had found this passage span very difficult to understand.
25-26	and the fourth day SOMETHING of the full moon was umm.. set? as a holiday.. fourth day of the full moon.. of the what? I can't think anything that comes here [INAUDIBLE] the fourth day of the full moon? was set (7secs) fourth day something of the full moon I DUNNO.. fourth day SOMETHING the full moon.. fourth day was set something as a holiday set aside? hmmm? holiday TO BE [A HOLIDAY].. I don't know about this here hmm.. set something it think it means like it was like kept as a holiday.. or like it was set as a holiday or something I.. set umm.. something was a holiday I think.. every fourth day of the full moon was.. there wasn't any event but then what shall I say?.. was set aside as a holiday the.. the one before [INAUDIBLE] the fourth day of the full moon fourth day comma umm so if it coincide with the full moon? no it's just too long it should be one word ah.. the fourth day um (8secs) no this one	RFL1 +Diff RFTL TLPara SWA WC Tr Und LAN Sou	Yasuko noted in her post-task interview that here too she had felt uneasy about repeating a word in such close proximity.

	I don't know.. can't think of anything .. I think there.. noun should be here.. on the fourth day SOMETHING of the full moon fourth day but [INAUDIBLE] where was it? full moon FULL MOON (6secs) day of the full moon no on the fourth day (8secs) day of the full moon but do you say of the full moon? FOURTH DAY OF THE FULL MOON? I understand the meaning but then do you say *day* of the full moon? night of the full moon?.. no because the whole day is a holiday so I think you say it's day of the full moon.. can you say day and then comes a day? hmmm.. doesn't sound right		
27	(5secs) and the SOMETHING last day.. and the SOMETHING and the last day all the victors (11secs) and the last day THAT'S NOT IT NO.. WHAT ELSE WOULD COME HERE?.. AND SOMTHING and last day should be adjective here? on the final.. and the final day is the last same as the last day so.. come on.. the [INAUDIBLE] and the.. I DON'T GET IT! SOMETHING and last day all the victors.. the day before? no the day before the last and the last day.. no because there should be only one word (5secs) the first no.. its not the first and last day because there won't be any victors on the first day umm.. I think just say on the last day..	WC +Diff Log/KO W GUP?	Yasuko was unable to say in her post-task interview whether or not she had in fact 'given up' on this item, but agreed that the blank could not be left unfilled.
28	all the victors *were* crowned because it's passive need [verb] to be..	Gra	
29	crowned with holy garlands of wild (6secs) wild flowers I think wild umm.. think it's.. it's actually the olive branches no? but then umm.. you don't say wild olives.. don't you? don't know.. but wild flowers looks okay..	KOW Col LK	Yasuko claimed to have been aware that the Olympic crown was made of olive leaves, but she apparently could not reconcile 'olives' with 'wild'. 'Wild flowers' just "came into [her] mind" as an "idiom" she knew.
30-32	so great SOMETHING the honour.. so great the honour so great was the honour that the winner of.. the foot race give the name to year of his victory (12secs) the foot race gave his name to [INAUDIBLE] the year is the .. the? names to (8secs) so great was the honour that the winner of the foot race gave his name to (7secs) SOMETHING to the year? next year? of his victory umm (10) so great was the honour	Gra TLPara +Diff	Yasuko noted in her post-task interview that she had found it hard to imagine that a year could be given an athlete's name. She then volunteered a parallel with the Japanese system of

because so SOMETHING was.. was that so SOMETHING that and then this so great umm..
this sentence formation I learned I think in school that the verb and then subject is reversed so the honour was great
but so great was the honour that the winner of the foot race gave his name to..
I don't know exactly what to.. what he gave his name to.. the year of his victory?
to the the .. why is year? I don't understand the meaning..
so the year of his victory I mean the year he won the race got this guy's name?..
doesn't make sense.. so everybody in Greece have to call the year like.. year of year of.. of somebody? STRANGE.. I don't know!.. I can't think of anything ..

Imperial eras.

Discussion of Yasuko's protocol data

A number of differences will be apparent between Claudia's protocol and that of Yasuko. Although Claudia makes limited use of running fillers (in her case, a characteristic sound '*hmhmhm*') in the early stages of her processing, these seem to fade out fairly quickly. Yasuko makes much more extensive use of running fillers, and these take the form of TL 'something' or, more frequently, its L1 equivalents '*naninani*' or '*nantoka*'. In her PTI Yasuko claimed not to have noticed using TL running fillers, but noted that the use of L1 fillers was something she had done since beginning to learn to read in her L1, when she had used these to stand in for *kanji* characters whose meanings she had not yet learned.

Another feature of Yasuko's protocol is the presence of rather longer periods of silence than were found on Claudia's transcript. There is a view (Maynard 1997) that the role of silence in interaction (and an element of at least imagined

interaction is plausibly inherent in think aloud) differs between Western and Japanese cultures, and that Japanese are much more accepting of even long pauses for thought. It cannot be ruled out that impacts on think aloud, too. The less 'structured' nature of Yasuko's verbal data is also conspicuous, as is the greater frequency with which indecipherable spans of data present themselves. (It should be noted that these are not practically open to review in post-task interviews.)

Apart from the use of L1 running fillers, above, a recurrent event we find in Yasuko's protocol is her concern (coded as 'KS') to mention that she acquired a particular phrase or rule in school English class. Although a similar elaboration occurs in GL1 data (typically mentioning that the informant read an item of information somewhere) JL informants commonly mentioned school English lessons. While it is true that until fairly recently most Japanese had fewer opportunities than Europeans to acquire English out of school, another factor may be illuminated by Yasuko's post-task answer to my question about why she had thought to mention *where* she had acquired certain bits of TL knowledge. Her suggestion was that some English phrases and structures given particular attention in school lessons ('SO [ADJ] WAS [NP] THAT..' would be a good example) might be especially easily recognised when encountered in text. The degree of rote memorisation ('over-learning') in Japanese school language education is comparatively high, and the acquisition of certain L2 items there may indeed be highly memorable. Yasuko mentions five L1 equivalents to TL items, and Claudia two.

While acknowledging that the distinction between access to syntactic knowledge and lexical knowledge (Ellis 1984, see also Read 2000:113) is far from cut and dried (and see the upcoming discussion of NCR data) Claudia appeared to refer to knowledge of grammatical rules in her processing of 12 or so deletions, and to access knowledge of phrases or collocations (where again, cf. Nattinger & DeCarrico 1992, no clear boundary obtains) in three. Yasuko seems to use grammatical knowledge in seven blanks, and knowledge of TL phrases in nine or ten. One clearly problematic aspect of TL structure for Yasuko is the use of articles. This is not surprising in that her L1 lacks an article system (The means in Japanese of indicating definiteness/indefiniteness do not readily map onto those of English.) and even otherwise highly proficient Japanese speakers of English may find article usage difficult. Yasuko's protocol contains some nine references to the word class of the missing item to Claudia's five or six. Whether this represents a step towards full recovery of the deleted item, a kind of mnemonic device, a way of salvaging 'face' before the microphone, or a combination of these motivates, is unclear. In her post-task interview Yasuko claimed that in classroom didactic cloze tasks, too, she would often identify the class of the missing word, and felt that this probably helped to narrow down the search for a filler. As an authentic part of her processing, she felt, identification of word class belonged in her think aloud report.

Just as we saw in Claudia's protocol, Yasuko's reporting seems to indicate processing almost entirely at a local level. Only in blanks (8), (21), and (38) does

she appear to have made use of passage information from out with the sentence context, and there is no indication that she activated information from beyond the paragraph. Three points must be kept in mind, however, when we seek to use think aloud data in regard to level of processing. The first is that the OLYMPICS passage does not appear to require much in the way of 'global' processing anyway: if we were to construct (and a degree of deliberate (re)construction might be needed) a natural cloze passage that extensively targeted the uptake of long-range information the picture might be different. Secondly, Alderson 2000 and others have suggested that by its very nature cloze encourages low-level processing. The regular mutilation of the passage by numbered blanks may indeed (and I was to some extent conscious of this in my initial self-as-subject processing of German passages) encourage a linear 'attention flow.' Moreover, comparative observations of (didactic) cloze task-takers at work and of individuals at work on classroom lexical inferencing tasks strongly suggested that the former shift their gaze significantly less than the latter. Finally, the cognitive demand of the think aloud task may in itself contribute to the 'localisation' of attention, for reference to longer-range passage content was, as we shall see, slightly less uncommon in NCR protocols. It could be argued that the lower cognitive load that NCR appears to pose may permit the task-taker to access content more widely.

Whereas Claudia only questioned the need for a filler word once, in item (26), Yasuko did so on in relation to items (35, (36) and (37). Given that Claudia did not recover deletions beyond (32), no real comparison can be made between the

two protocols in this respect, however; the items which Yasuko queried are quite clearly more plausibly left unfilled than is (26). Given that the requirements of the cloze task (one filler word each time, *every* time) were spelled out very clearly before the session, informants' questioning of the need for a filler is perhaps best seen as a way of demonstrating their understanding that the parent span of passage also makes sense just as it is. I was struck by Yasuko's unusually strong suggestion that in item (37):

"..no there shouldn't.. there cannot be anything here.."

and took up the issue in her post-task interview, which was separated from the verbal report session only by a short break. Her remark struck me as odd, too, because she had already contemplated, only to immediately reject for reasons not made clear in the think aloud, the filler 'feel (worthwhile). When I asked why she had rejected 'feel' as a filler Yasuko very quickly replied (and this suggests that her response was authentic, i.e. based on recollection rather than on a constructed rationalisation) that it was not appropriate because the sentence would then carry the connotation that the training had not in fact been worthwhile; to her, it seemed, 'feel worthwhile' bore the implication or connotation of an illusory state. This, she felt, was far from the intent or message of the passage, and so she rejected 'feel' as a candidate filler. The rich rewards and public honour mentioned in the paragraph, and the suggestion (item (39)) that runners were willing to risk dropping dead at the winning post, may have contributed to what appeared to be a firm understanding on Yasuko's part that the final paragraph implied a positive

view of the athletes' dedication and discipline. In other words, her statement, above, may convey not just a realisation that no word is syntactically *necessary* here, but rather that for optimal coherence (within her interpretation) there ought *not* to be a filler in the first place. This small event is interesting, then, not so much for the strength with which Yasuko phrased her suggestion (or even its misguidedness in a cloze context) but rather because it may indicate a 'veiled' higher level of interpretation and integration than the protocol superficially conveys.

By the same criterion of filler actually entered, Yasuko's SEMAC score on the first 32 deletions of the OLYMPICS passage was 66% (exactly the same as Claudia's score) and 71% on the 41-deletion version. Her time-on-task was approximately 34 minutes.

6.3 The less productive thinker aloud: Harumi

The importance (or, perhaps, the 'urgency') of retrospective data depends in part on the clarity of the original think aloud reporting. Claudia's protocol, above, was on the whole quite readily comprehensible, but greater importance may attach to the post-task interviews of informants like JL1 Harumi, whose protocol is rather less interpretable. In these cases, the informant may have to confirm or correct the researcher's understanding, or in an extreme case even to clarify her actual words. The think aloud performance of JL1 informant Harumi does not place her in the 'low verbaliser' category, but although the quantity of thinking aloud is there, it is weak in terms of the insight it allows into her processing behaviour. As will be

clear from the following chapters of this paper, I have found think aloud to be an unpredictable and inconsistent source of data, and Harumi's case is one example of less successful reporting. Some of the blame for this lies with me, for I was unable to audit Harumi's reporting adequately, and circumstances made it impossible to schedule a post-task interview that was *authentically* post-task. The fact that Harumi's reporting session had to be scheduled in the early evening may have been a factor in her poor performance. That said, these things happen, and the longer and more requiring of supervision the data-gathering procedure is, the greater the scope it contains for disappointment.

Harumi was recruited into the pool of potential informants through a friend who had already volunteered. As assessed by her placement test scores in reading comprehension and writing (though not in listening comprehension) Harumi displayed average English ability for her year group of course peers. In classroom activities and in a semester test of the main skill areas, she repeatedly scored within a few points of the average, and her level of application to–and interaction in–classroom tasks was entirely adequate. In a short trial-cum-orientation think-aloud task, moreover, she reported well enough to be included in the informant pool. It was not until I audited Harumi's thinking aloud on the OLYMPICS data-elicitation passage (for I had not been able to observe her think aloud session throughout) that I noticed her fairly idiosyncratic processing behaviour.

Harumi essentially makes two 'passes' at the passage, with some attempts at filling blanks in each. It was fairly common (approximately one third audibly took this step) for informants to 'scan read' part of the passage (or, in a very few instances, apparently to scan more or less the entire passage) before beginning to work seriously on the task of filling blanks, but few effectively processed the passage twice, as Harumi did. Invited to talk about her normal way of approaching test passages, Harumi confirmed that she "pretty much (always)" carried out a second run-through of a task in order to check her earlier conclusions. She added remarks which I interpreted to suggest an awareness that answers might be inconsistent and thus require revision, and that information later in the text might help answer early questions. On the face of it, spending time on filling deletions on a first pass seems to be a less efficient strategy than one involving a first reading followed by recovery, but it is plausible that Harumi hoped to fill easier blanks during her first pass (cf. Cohen 1998:230ff for an account of what may be an analogous approach to the cloze task by a Portugese L1 informant.) Harumi had clearly encountered 'gap' tests of some kind in her school language education (and as she also studied French at college level she had quite possibly encountered natural or 'rational' cloze in that language) but my inference from her remarks was that she had been taught the technique of in general having two 'tries' at foreign language text-based tasks. Other Japanese of my acquaintance have told me that Japanese school teachers of English quite often recommend such an approach to their pupils.

Again, I show Harumi's protocol content, derived codings, and post-task interview insights. (For scheduling reasons Harumi's interview took place two days after her think aloud session.)

THINK ALOUD PROTOCOL JL1 Harumi 'FIRST PASS'

Del.	Protocol extracts	Code	Comments
1	the Olympic athletic festival held.. held every mmm.. athletic festival held every *four* years	???	
2	eventually.. became its local character.. local character?	???	
3-4 (4-3)	became first a.. event and then.. after the rules *against* foreign competitors had been waived ..international... first a *national* event?	Col?	Harumi confirmed post-task that 'international' had made her think of 'national', so that 'Col' is the appropriate coding of TA data.
5-8	..one knows exactly how far back.. WHAT IS IT? [6secs] GO BACK.. some official.. mmm? ..one knows exactly how far back Olympic go [bell sounds] SIX O' CLOCK? [THAT'S RIGHT] how far back the... mmm? how far back EH... one knows the games games [INAUDIBLE] plain mount Olympus IS IT THE MEETING-PLACE?	???	??? *Harumi's apparent concentration on deletion 5 is broken by the bell.*
9	plain *of* mount.. EH mount Olympus? EH... mount mount.. EH THAT'S WRONG..	???	This is in fact an acceptable filler, Harumi could not explain why she appeared to reject it here *only to then enter it as a filler.*
10-11	many thousands of spectators *came* from all part of Greece but mmm... people married woman was.. was admitted.. [INAUDIBLE] people Greece.. Greek.. mmm? I UNDERSTAND..	???	*(Researcher has to leave room at around this point. Harumi is asked to continue thinking aloud.)*
12-13	[INAUDIBLE] slaves women and dishonoured person [sic] *were* not allowed to compete..	???	*In reading aloud, Harumi appears to disregard the plural ending of 'persons'.* In her post-task interview, she noted that the plural of 'person' is 'people.'
14	the exact SOMETHING of events is uncertain ..exact *purpose* of.. mmm?	RFL1 ???	Harumi could not shed light on the juxtaposition of 'purpose' and 'events' here.

15-17	but events.. boys' gymnastics.. horse-racing field events /diskasu/ WHAT'S THAT? throwing and.. WHAT very important foot .. exact kind of events is uncertain but events WHAT WILL FIT? [5secs] THIS IS HARD! boys'.. boys' gymnastics.. horse racing field events as [INAUDIBLE]	UPI +Diff	
18	There was boxing and wrestling and special tests.. varied ability such as the pen.. WHAT'S THAT? I DON'T UNDERSTAND [INAUDIBLE] mmm AND mm..	UPI	*(Researcher returns to room.)*
22-24	evening of the third day... FULL MOON THAT'S NOT IT [INAUDIBLE] sacrificial offerings to the heroes of the day ...	??? Phr	Harumi's post-task comments indicate that she was familiar with the phrase 'x of the day'
25-29	was set... set on the THIS IS IT? first NO, WRONG.. day all victors were crowned with holy garlands of wild SOMETHING OR OTHER from the sacred wood [INAUDIBLE] SAKAKI ... SAKAKI SAKAKI.. NOPE IT'S NOT [INAUDIBLE] SOMETHING YOU WEAR AROUND THE HEAD [GESTURE OF CIRCLING HEAD] and the fourth day... mmm AND full [INAUDIBLE] date.. day day day NOPE WHAT IS THAT THING TENNIS CHAMPIONS AND PEOPLE LIKE THAT RECEIVE?	L1Eq? UND KOW	*Harumi's gestures indicate she knows what the victor's garland looks like*; *'sakaki'* is *Eurya ochnacea*, a tree species sacred to Japan's *Shinto* religion and whose leaves loosely resemble those of laurel.
30-32	so great the honour was so great was the honour.. the honour was so great that.. that [INAUDIBLE] the winner of important foot races give his name to.. WHAT IS THIS? year of his victory..[TRANSLATES SPAN]	TLPar Tr	

SECOND PASS

1	in ancient Greece athletic festivals are very important had strong religious associations.. the Olympian athletic festival held every every mm.. every *four* years RIGHT? NOW IT'S FOUR YEARS IN BETWEEN RIGHT? IN THOSE DAYS?	KOW	*This verbatim extract shows how Harumi appears to pay little attention to surface text features such as tense.* In her post-task interview, however, when shown the typed extract "In ancient Greece athletic festivals.... strong religious associations" and invited to complete it without reference to the passage, Harumi supplied the correct tenses without difficulty. In her post-task interview, Harumi suggested that 'four' was a logical choice given that the current Olympic interval is also four years. Her apparent hesitation here is interesting in view of the fact that she had *already filled* the blank. Harumi could not comment this point.

2	in honour of the... eventually become its local character	???	*Again, Harumi chooses 'become' but this time uses present tense.* In her post-task interview, Harumi could not recall whether, as it appears from the protocol, she was positing *became* as a filler for deletion (2).
3	became first national event [INAUDIBLE]	???	(See below)
4	and then after rules against foreign competitors have been waive waived .. international	??? Col	*Again, H. appears to mix tenses*, but was able to supply the correct past perfect in response to a 'blanked' prompt as above. *She also seems to hesitate despite having already filled in the blank.* Asked in her post-task interview if she might have been 'checking' her earlier filler, Harumi seemed unsure. She cited the collocational link between *rules* and *against*, and claimed that when she saw the word 'international' she could infer 'national' for deletion(3): "These go together."
5-6	*no* one knows exactly how far back Olympic games can go back HOW MANY YEARS AGO can [INAUDIBLE]	???	(Researcher has to leave room briefly.) In her post-task interview Harumi claimed the filler "just came to mind".
7	but official some official mmm? date date from SINCE SEVEN HUNDRED AND SEVENTY SIX BC SEVEN HUNDRED AND SEVENTY SIX BC some official SOMETHING official official today.. [INAUDIBLE] one exactly mmmm Olympic [INAUDIBLE] by some official.....official mmm official WHAT? official [politicians] um.. pretty [hard]	Tr +Diff	In her post-task interview, Harumi confirmed that she had glossed this extract in her L1. She could not account for the extraneous 'today one'. Harumi could not confirm the barely audible utterance glossed here as 'politicians'.
8	that games games mmm held mm? the games were held EH.. WRONG? THE PLACE games games SOMETHING place games had... *had* a place in August *take* place YES? *took* place... the games *took* place in August	Sou	In her post-task interview Harumi confirmed that she had sounded out various candidate fillers and chosen one which "sounded okay." Told in her post-task interview that her filler was in fact correct, Harumi could not explain why she had thought it wrong.
9	on the plain plain mount Olympus mount umm mountain [INAUDIBLE] mountain mm? on the plain on the plain in mount Olympus THAT'LL DO...LIKE THAT	???	Harumi filled this deletion on her first pass, but then appeared to reject the filler without, however, changing it. *(Researcher returns)*

10-13	but man... woman EH? was.. SO IT'S [many] ISN'T IT? many married woman was admitted even as a spectator... slaves.. slaves women and dishonoured persons *were* not allowed to compete [INAUDIBLE] WHAT CAN I FIND? old old [INAUDIBLE] wrong /grek/ as even as a [...] greece greek married woman was admitted even as mm... [A SIGHTSEER].. slaves women and dishonoured persons were not allowed to compete..	???	In her post-task interview, Harumi was unable to clarify the sequence 'people married woman' in her first pass. She was also unsure whether she had considered 'Greek' as a potential filler for deletion 11. It was unclear from the tape whether Harumi was saying 'man' or (in an idiosyncratic pronunciation) 'many' the second time, and she could not elucidate. Curiously, Harumi also had difficulty in confirming the barely audible L1 item I have glossed as sightseer (*kembutsu-nin*)
14	the exact *kind* of events? [INAUDIBLE]	??	In her post-task interview, Harumi suggested that 'kinds' would be grammatically more accurate.
15-17	but events WHAT IS THIS? boys' gymnastics events like boys' gymnastics horse racing field events mmm.. SOMETHING events field events um... RECREATIONAL EVENTS? track and field OR field events mmm throwing umm track and field THIS IS ALL MUDDLED UP and field events [INAUDIBLE] field events...very important foot races.. the exact.. but events umm.. were were? STRANGE THOUGH.. horse racing field events events *such* as [INAUDIBLE] discus and [INAUDIBLE]? DISCUS? and and and eh? AND other... other very important foot races	RF +Diff	In her post-task interview, Harumi claimed to be unsure about which events 'track and field' (a term imported into Japanese) includes. Harumi asked whether 'events *were* X, Y and Z' was grammatically acceptable. I noted that it was not on the list of SEMAC fillers. She also confirmed her gloss of the meaning of 'discus' (*enbannage* in Japanese.)
18-21	[INAUDIBLE] there was many tests special tests depending on.. mm? [INAUDIBLE] and... EH?.. winner of which excelled in running running horseracing gymnastics running running AND WHAT? discus [INAUDIBLE] umm.. racing race discus and javelin throwing and racing? WHAT CAN IT BE? and no THAT'S NOT IT...became? there was [INAUDIBLE]	UPI SWA	*Note the misreading of the actual passage*. In her post-task interview, Harumi claimed to have been aware of the single word filler rule, and to have been looking for such a word with the meaning of 'depending on' In her post-task interview, Harumi confirmed that she did not know the meaning of 'pentathlon.'
22-24	and evening of the third day was.. was to offerings.. the evening [INAUDIBLE] third day was.. was determined EH? to umm set /seto/ /seto/ /seto/.. umm *set* to sacrificial offerings..[INAUDIBLE]	**SSf** Tr	Harumi claimed to have noticed 'set...as a holy day' later in the sentence, and felt this might have influenced her choice of filler in (23). She also claimed to have followed this by glossing 'heroes of the day' into her L1 even though she had already filled (24) on the first pass.

25-26	and the fourth day evening of [INAUDIBLE] was set set to [INAUDIBLE] STRANGE..	CL?	*The barely audible final remark here could be interpreted as implying Harumi's own lack of confidence in her comprehension of this span.*
27	on the.. on the.. on the SOMETHING last day	RFL1	In her post-task interview, Harumi confirmed that she had repeated the phrase to "see if something came to mind." (See below)
28-29	on the... with holy garlands of wild [INAUDIBLE] WHAT IS IT? [GESTURE]	KOW	Harumi accurately- sketches an Olympic crown in the margin. In her post-task interview, and confirmed that she had glossed 'sacred wood' as 'material from a sacred tree', noting the similar appearance of *sasaki* and Olympic garlands.
30-32	OKAY.. his name the honour.. that the winner of bababa gave his name to... the honour that the winner of bababa gave his name to... I CAN'T IMAGINE... [INAUDIBLE] [SO YOU'VE GOT THEM ALL NOW.. FINISHED?] uh-huh.. FINISHED	??? RF	*Harumi suddenly switches to 'sound' running fillers.*

Harumi's first pass, especially, features the reading aloud (often at high speed) of spans of passage which may be 'telescoped' into one another. At some stages it is impossible to separate her processing of individual deletions, and so these must be shown together. As shown by the frequent recourse to [INAUDIBLE], I experienced considerable difficulty in deciphering some of Harumi's comments in Japanese. This puzzled me until I realised that she was at times using a regional dialect of Japanese which she did not use in lessons—a feature unwelcome in this context, but one which nonetheless implied a highly authentic interaction with the task. I had been aware that Harumi's family came from a rural area of southern Japan, but her classroom interaction with others was regularly conducted in standard Japanese.

The relative poverty of Harumi's protocol illustrates that even an apparently promising informant may fail to produce data that allows much insight into her processing. Having by this stage elicited (or, rather, attempted to elicit) think aloud data from several JL1 low-verbalisers, I was well attuned to the need to avoid any indication that the task had not been performed in a satisfactory manner. As a general rule in post-task interviews with Japanese informants, I put prompt questions in English as much as possible, using Japanese only when clarification was asked for or needed. The Japanese language possesses a wide variety of ways to 'soften' the tone of questions (*yasashiku-kiku*) or phrase suggestions, etc. in a variety of 'comfortable' ways, and I did not feel confident in my ability to avoid an inappropriate tone. In the event, Harumi appeared to be aware of the limitations of her own reporting. Although at a few points in her post task interview she appeared reluctant to respond to my questions, at others she seemed genuinely *unable* to recall aspects of her processing, or even what (I had expected would be) fairly salient events therein. Harumi's 'spontaneous' remarks were conspicuously few, and brief, and virtually all of the comments shown below were in response to my prompt questions.

This is not to imply that I have any less confidence in Harumi's retrospective comments than in those of, say, Claudia, above The fact that Harumi repeatedly claimed not to recall certain details of her processing behaviour may rather suggest that the information she did volunteer is limited to what she believed to be accurate.

I will not discuss Harumi's protocol in further detail, except to repeat that I had no reasonable grounds to anticipate that she would perform so poorly on the task (Recall that I was perpetually short of informants, and could not afford to turn any away without good reason—particularly in view of my snowball/invitation-by-friends recruitment method.) and in particular so often so often articulate unclearly. It was doubly unfortunate that I could not schedule a post-task interview with Harumi until two days later, as she might earlier have been able to elaborate on her processing. On a number of occasions, as in deletions (8) and (18)–(21) she rejected an unstated idea or filler. Again, had I been able to audit her reporting without distraction I might have been able to interrupt her with a request to clarify these events. I did ask Harumi, post-task, if she could recall what she had been saying or doing during the uninterpretable spans of reporting, and her answer suggested that these intervals had contained "strange" (*hen*) ideas in which she had little confidence. (A German informant, Wilfried, had made a similar comment in his (immediately) post-task interview, suggesting that some of his thoughts had been so "mixed up" (*durcheinander*) that they would have made little sense.) How deliberate or even conscious inaudible reporting may be is far from clear, but it appears that informants can readily come up with some kind of *post facto* rationalisation. One new insight may have emerged from Harumi's post-task interview, however, and this is taken up next.

A small insight?

An unexpectedly detailed insight into a fairly minor aspect of Harumi's processing repertoire came when I asked her to tell me more (as we audited the relevant tape extract on a transcription device allowing automatic repetition of 2–4 second extracts.) about her second-pass processing of deletion (27) I had assumed that the verbalisation I had heard as

"..on the.. on the.. on the SOMETHING [NANTOKA] last day.."

reflected the not unusual tactic of repetition in the anticipation of a potential filler coming to mind. Harumi confirmed that this was "pretty much" what she had been doing. After two or three reviews, however, she remarked in passing on a feature which I had only superficially registered, namely that the first two repetitions lacked the L1 running filler 'nantoka' (SOMETHING) while the last included it. I asked if this was important or significant, and I inferred from Harumi's reply that in the first two instances she had been, as it were, holding an aural or mental space 'open' as she waited for a filler to appear. When none did so, she added the running filler SOMETHING on the next repetition. The presence or absence of a running filler, then, may imply something about the informant's perception at that moment of how close she is to filling the blank. Apparently analogous sequences of 'open space' followed by running filler (The use of L1 items, nonce words, or sounds as fillers appears to reflect an individual though by no means always stable preference, as seen in Harumi's final item.) has been noted in the protocols of others, but none of these was able (if even invited) to elaborate on what this might mean in her case. Investigated while still fairly fresh in the minds of a sufficient

number of informants, this might form an interesting avenue for further research.

6.4 Pair-condition think aloud informants

I turn now to an illustration of pair-condition think aloud (PCTA) protocol data, some arguments for this format were rehearsed in section 4.14, but certain issues remain to be discussed. First of all, it will be clear that paired condition reporting potentially greatly complicates the data elicited, for this will inevitably add to the (already complex enough) interaction of informant and text that of informant and partner. The affective dimension of this interaction may reflect age and gender factors, perceived divergences in language ability or other bases of status ranking, as well as degrees of assertiveness, and task-commitment. Any of these may impact on the mutual interest and/or respect that informants must have if the elicitation task is to be productive, and thus on what gets reported and by whom. With regard to the interactions of JL1 informants especially, perhaps only a native of that culture can fully understand the intricacies of what occurs. Haastrup 1991 notes that pair informants do not always share intuitions, cues and other information with one another, and this is a point I return to below.

Volume of data is a more practical concern. PCTA protocols can be very long indeed, and given the conversational dimension to the reporting format we should not be surprised that by no means all the content appears directly relevant to the recovery task. I have thus taken the liberty of editing these protocols for presentation here, and, further, I have 'condensed' some exchanges into what I took to be their essentials, as in [X queries Y's confidence in the filler] or [Y

defers to X. A number of the codes listed in appendix 2 pertain solely to pair-condition reporting, and these are seen as the minimum needed to categorise the recurrent events noted in pair-reporting. Most of these will need little explanation in concept, although it should be noted that the code 'Did(actic)' needs to be interpreted broadly. Many pair-condition informants appeared to consciously avoid 'talking down' to a partner, so that when recovery-relevant information had to be communicated this was often done in terms of a question that pointed it up more or less unmistakably ("COULD THIS BE ANOTHER KIND OF SPORT... LIKE LONGJUMPING?") or in the form of a 'face saving' gambit such as ("YOU PROBABLY ALREADY THOUGHT OF */rekoodo/*, RIGHT?".) Deference ('Defr') of one partner may also be a sensitive area. I interpreted from the exchange below that B had deferred to A's opinion:

A "well, I'm almost sure it's *laurel*...why do you say [it's] wood?"
B "hmmm...SACRED WOOD.. FROM SACRED WOOD.."
A "no no.. WOOD is like *forest*.. you see? the tree.. I mean it's about trees [...] IN A SACRED FOREST.. it's something like this [GESTURES]"
B "OH..SO [MIMICS A's GESTURES] like..like a crown of..umm small leaves"
A "EXACTLY!"
B "WELL OKAY.. laurel?"
A "hmmm..that's laurel I think"
B "yeah? okay" [INDICATES THAT A. SHOULD WRITE IN FILLER]

Post-task, however (informants were interviewed separately on this point) neither the deferrer nor the partner deferred to seemed very willing to explicitly confirm my inference. This was puzzling, for while one might anticipate that the deferring partner might wish to downplay her behaviour, the one deferred to would seem to

have little to lose by admitting that her opinion carried the moment. In regard to the session extract above, it is worth mentioning my inference at the time that B's gestured invitation to A to write in the filler constituted an implicit rejection of A's conclusion. Both partners had pens in their hands, and had fairly consistently been taking turns at writing in fillers. I cannot be sure that it was B's turn to write the filler, but A seemed slightly surprised at being invited to do so herself. Her L1 comment as she did so, "*Na gut*", had, for me, the connotation of "Have it your way." It could be argued that by this gesture B. was downplaying her deference to A's quite forceful opinion about passage meaning and choice of filler. As my GL1 think aloud informants were very reluctant to be video recorded, fairly subtle gestures like this could easily have been missed, and could not readily have been recovered from the audio-protocol alone.

6.5 GL1 Protocol data from GL1 (Anneke & Fred) and JL1 (Mitsuo & Arisa) pair informants

I look now at the interaction of one GL1 informant pair: Anneke & Fred. These two (whose protocol was selected on no other basis than having been at the top of the pile) were recruited from the same intact course group, and were felt by myself and another of their teachers to be fairly closely comparable in terms of language ability; their scores on classroom paper activities were broadly similar. Fred hailed from the south of Germany, and Anneke from the centre of the country. Fred's outwardly very diffident manner had appeared to irritate a few other individuals in his class, but he and Anneke seemed to get along quite well and had in fact effectively self-selected as partners for pair-condition reporting.

Despite (?) episodes of what appeared to be mutual teasing Anneke and Fred seemed to cooperate quite well during the task session, and appeared to be in no hurry to finish. On at least two occasions one gestured to his or her watch, and the other replied with a gesture which seemed to indicate that time was not a factor. The pair agreed before hand to report mainly in their L1. Extracts from the PCTA protocol of Anneke (A) and Fred (F) are shown below with the informants' verbalisations separated according to the deletion being processed at the time. The layout of this pair-protocol has been adapted for clarity. Informants' utterances are not (for software reasons) positioned precisely enough to show the exact moments of speech onset and duration of overlaps, as is sometimes done in the short extracts cited in discourse analyses. I do not believe this detracts seriously from the information the protocol provides.

[NB in the extract below regular type indicates L1 verbalisation, with TL in italics.]

Del	Protocol extracts	Codes	Comments
1	F. umm probably the number of... how often or not it was celebrated.. A. umm.. so every *two* years.. F. no that's not it! it's *four* isn't it? now it's four.. A. oh right [LAUGHS] F. but I think that in earlier times the intervals were probably even bigger.. well let's write *four* four *four* shall *I* write [it]? A. okay	KOW Diff CP	*F. smilingly but firmly rejects A's suggested filler. A. seems embarrassed. F. waits for A's approval (a nod) before taking up the pen and inserting the filler.*

2-3	F. umm I think here that um...at some stage it umm...its local character A. maybe *lost*? F. *abandoned* or *lost* A. so.. lost? F. *lost*, yeah.. A. became first... a *national* F. umm.. isn't it *inter*national? A. well wasn't it in Greece at first? F. oh right...something like over the whole country...is national the whole country? or should it be *nation-wide* or something like that? A. well.. *national* or *nation-wide*? F. I dunno if national and nation-wide are the same so maybe [...] A. and after the rules against foreign competitors had been [...] okay.. *no* one knows knows exactly how far back.. the year.. *the* games go.. F. ummm yeah.. yeah I think perhaps it's more likely to be the traditional olympic games... now there'll be a distinction made between now and classical times.. A. perhaps those of today or something [...] perhaps *modern* olympic games go? F. no no I'm pretty sure that.. as you say.. it refers to everything.. just *the* olympic games right? A. {NODS} those of today?... oh..okay..	CBA CP KOW CP Solo CBA	*A. appears irritated by F's indecision, quite abruptly asks him to choose 'national' or 'nation-wide'. A writes in 'national'. She then writes in ' in against' in (4) without discussing this with F. [F. SMILES, SHAKES HEAD]* *A. carries on to fill (5) and (6), but then questions and amends her own suggestion for (6) F. argues that her first suggestion was correct. A. appears to misunderstand but accept F's argument (She seems to think he is agreeing with her interpretation that only the modern games are being discussed here.) but carries on to the next deletion.* In the post-task interview (conducted at the informants' express request, jointly) A. admitted that she was annoyed at F's reluctance to choose between two items A. felt were "much the same". F. said he had noted A's annoyance. A. confirmed that she thought F. had accepted her interpretation. F's remarks suggest he was aware of this but preferred to "just leave it" and move on.

The extract above illustrates how pair-condition reporting lends itself to readier interpretation than solo think aloud, as each informant's need to make herself understood to her partner means that she has to structure her verbalisation in a readily understandable way. At the same time, many small (and for the study of cloze processing, perhaps incidental) insights into informant's grasp of lexical meaning present themselves such as Fred's statement of his expectation of what is to come next or soon in the passage: "...now there'll be a distinction made between now and classical times.." This kind of explicit anticipation is extremely rare in solo think aloud data, and in this instance may be taken as evidence that Fred has not read far ahead of the item currently in focus. While the paired reporting task may makes it difficult for an individual informant to 'read ahead' of her partner, this is not unknown. The exchange below suggests hat one partner, Kaori, has in fact been reading ahead even before proposing to her partner that they do so:

Kaori: "[...] yeah but if we.. we could look at what the next part says and.."
Noriko: "both of us?.."
Kaori: "yeah.. maybe we can find the word.."
Noriko: "so we want to find another kind of sport, right?"
Kaori: "I guess so.. but there's nothing.. did you see anything?"
Noriko: "Haven't [read it] yet.."

Anneke and Fred, however, appear to focus quite tightly on the local context of each deletion. The extract below contains a clear instance of Fred correcting his partner's interpretation by reference to textual evidence that she had not yet noticed: the logical inference that "No one knows exactly.." is incompatible with

Anneke's apparent interpretation of "some official...." as implying a human agent. Why Fred appears to correct Anneke's accurate reading of the passage in the fifth turn, below, is unclear here and was not clarified post-task

[NB in the extract below regular type indicates L1 verbalisation, with TL in italics.]

7	A. the olympic games go.. but some official.. official who? people who've looked into it? F. Really? A. well perhaps historical researchers or.. F. yeah.. or it could be historical dates too A. or *historians* F. well.. but then the clause before isn't right A. ah, you're right [INAUDIBLE] F. if [we put] historians here and [there we have] date from...historical dates, right? A. but date is a *verb* here isn't it? so they date from.. F. no.. are dated from.. A. oh... F. so they could be just numbers..so let's say but some umm official *numbers* date from seventeen..seven hundred and seventy six umm since seven hundred and seventy six umm..right? A. okay..let's do that.. F. umm...some official [...] *evidence* or something.. A. or *evidence* (*Beweise*) F. but that doesn't fit does it? [LAUGHS] naah.. let's just do the next one A. [LAUGHS, SHRUGS]	CP Tr Gra CP LFN Ah!	*F. and A. move slightly closer together. Eye contact becomes more frequent. At one point A's lips begin to move but she produces no audible verbalisation that F. reacts to.* *A. appears to have to understood 'date' as an active verb form, while F. seems to interpret it as passive. (GL1 would require a passive construction in this context.)* *F. produces a filler ('numbers') to which A. agrees, but F. then offers an alternative ('evidence') following which A. suggests the L1 equivalent.* Post-task, A. felt that she and F. had their pair processing of this deletion had "gone really well." A. claimed that she had been translating during the ca. 6 second interval of silent speech. F. confirmed that he had also glossed the extract in the L1. A.: "But [F.] does it better.", both laugh. Asked about how they decided to move onto the next deletion, A. gently chided F. for giving up too easily. F. replied that he had run out of ideas for this item, and had wanted to move on. A., he added, hadn't seemed to object. A. had 'gone along' with F. slightly unwillingly.

JL1 pair-informants Mitsuo.& Arisa

I show below an extract from the PCTA protocol of two JL1 informants also recruited from an intact course group (The full protocol is shown as appendix 5.) Again, placement test and class test scores, and the assessments of two or more teachers placed these informants at comparable levels of effective proficiency. Mitsuo was noted to have a wider than average vocabulary knowledge and reading fluency, while Arisa's listening comprehension was superior. It was also noted in teachers' comments that despite a lower level of confidence in his own ability, Mitsuo was more willing to (in one teacher's words) 'take chances' and make the most of the English he knew. Both informants had acquired their knowledge of English very largely—or in Mitsuo's case entirely–within Japan. Arisa had spent a period of some months in the USA as a young teenager, and may have retained some out-of-class contact with native speakers of English. In an orientation session, Mitsuo had thought aloud without much apparent difficulty, whereas Arisa had been considerably quieter despite scoring well on the orientation passage. Although the two to some extent self-selected as partners, I was apprehensive that Mitsuo's contribution might overwhelm Arisa's, especially as it was clear that Mitsuo preferred to report in Japanese. Arisa appeared content to go along with Mitsuo's decision.

[NB in the extract below regular type indicates L1 verbalisation, with TL in italics.]

Del	Protocol extracts	Codes	Comments
1	A. *Four..* M. four *four* years.. A. let's write it.. M. yeah write it.. [A. READS AT LOW VOLUME] M. let's read on... we don't know what's written.. A. [LAUGHS]	KOW Rprt.	In their post-task interview, A. and M. confirmed that they "already knew" the 4 year interval. M. claimed that he had intended to read "half" of the passage before beginning to fill blanks, but A. had filled (1) almost at once. Note that M.'s next focus is on item (7).
7	M. what's that official?... a document? A. I don't know M. official DOCUMENT (*koshikibunsho*) goes back from 1976 to..	Col CP Tr	In the post-task interview, M. cited the link between official and documents, papers, etc. : "You see [the words] together." M. also confirmed his misreading of the date, and agreed that at this stage he thought the modern Olympics were the topic here.
5	A. mm.. M. perhaps it says when it started.. [READS AT LOW VOLUME, THEN APPEARS TO TRANSLATE] [...] *no* one knows' is right..*no*? A: mmm.. M. nobody knows.. but.. seven hundred and sixty years... A. uh.. [LAUGHS] M. perhaps, because it's *some* A. [WRITES] *no* one knows... let's write in pencil what we think.. later on... M. let's make this one *no*..	Tr CP Diff	In the post-task interview, M. confirmed that he had been translating the passage spans he had read, but could not recall the clues that led him to the filler. *M. reads date correctly.* A. claimed in post-task interview that she had laughed because she thought M. had realised his earlier error. She confirmed that she suggested writing in pencil (This was allowed in TA) because she anticipated the need to revise fillers. *M. appears to suggest a filler for an already completed item.*

6-8	M. something comes here, but what? A. the olympic games [M., A. LAUGH] A. docu.. M. DOCUMENT (*kiroku*) so... record? document? [A. FAILS TO RESPOND TO M's QUESTION] ah... *took* place in August.. is right A. mm.. [READS AT LOW VOLUME] M: mount.. mount.. ah there's mount olympus	Gra CBA CP Phr	Only when I confirmed (during the post-task interview) that 'documents' was a SEMAC filler in (7) did A., then M., notice that a plural was required. M. confirmed that he had known the phrase 'take place'.
9	A. mm? M. plain.. A. [LAUGHS] *of*?... don't know.. M. shall we make it *of*? mount olympus.. no.. *took* place.. A. [WRITES] took place [INAUDIBLE]	JSR CP Cha	In the post-task interview, neither could explain how they arrived at the filler for (9), only that it seemed acceptable.
13	M. this is *were* not allowed to compete.. A. eh? M. slaves and women were not citizens.. A. were not.. M. were not.. they couldn't [even] attend..	Gra KOW	In the post-task interview, M. claimed to have filled (13) on the basis of a grammatical need for an auxiliary verb. He also claimed to have known that only males could hold citizenship in ancient Greece.

The extract above demonstrates clearly that pair-condition think aloud does not always feature roughly equal input from both partners (or indeed to strict adherence to consecutive item order) for Mitsuo appears to have done the bulk of the work of recovery. At times Arisa appears to be either re-reading the current span of passage, or reading ahead on her own. At one stage she fails to respond to what was observed to have been a clear request from Mitsuo for confirmation

and/or agreement. Mitsuo apparently tries to process deletions (6)—(8) together, which may be a sign that he felt he could, or had to, work at his own pace.

Like Anneke and Fred, above, Mitsuo and Arisa chose to be interviewed together post-task. Mitsuo made no adverse comment in that session, but when I later interviewed him alone he admitted that he had been at first been rather disappointed with his partner's passive attitude (which he was aware was not due lack of TL ability) and he confirmed that he had felt very much as though he was working on his own in the earlier stages of the task. After a time, however, Arisa seemed to become more involved, and it appeared that she had felt the need to read a good part of the passage carefully before she could make a contribution. While it would be an exaggeration to say that A. and M. made two passes at the passage task, they seemed unconcerned at passing over several deletions without discussion, and at (in Arisa's word) "skipping" others which could not be filled in short order. I waited to see if they would return to these, and in the course of processing deletion (24) they are heard to agree to review their efforts so far. This exchange implies an appreciation of the potential value of earlier recoveries to deletions encountered later on, and reveals one partner's awareness of the risk involved in going back over the passage:

A. isn't it 'of'? [LAUGHS]
M. as the previous spaces are empty I'm uncertain...
A. let's read the previous parts again..
M. yeah let's.. but we may get stuck in the middle..

The pair then return to the very beginning of the passage, paying attention mainly to blanks left unfilled:

1,	M.	Tr	In the post-task interview, M. was
3,	AS THE EVENT WAS HELD	Gra	unable to recall whether he had
4,	(*okonawarete*).. became	CP	been, as it were, re-opening deletion
6,	something	UPI	(1). A. appeared to work on (3), while
	A.	Gra	M's next focus is on (4). A. and M.
	became first..	CBA	were unsure whether they had been
	M.		operating separately at this stage, or
	here it is.. subject.. so *of*? *of*		in tandem or in 'leapfrog' mode.
	foreign competitors have been		
	/waivd/.. is *of* correct?		
	A.		
	what's this?		
	M.	UPI	M. claimed to have been thinking of
	this I don't understand..		the L1 genitive construction 'X *no* Y'
	/waivd/..		(the X of Y/X's Y). He also said he no
	A. [LAUGHS]		idea of the meaning of 'waived'. A.
	M.		said she had laughed at M's
	this must be *of* of foreign		pronunciation of the word (which she
	competitors... back to [...] no		could pronounce correctly) but also
	one knows.. what is it? WHAT		at her own uncertainty as to its
	COULD BE PUT IN HERE? go back..		meaning. M. appeared to take no
	there's go..		umbrage at A's amusement.
	A.		
	not *the*?		A. claimed to have filled (6) on the
	M.		basis that an article was "perhaps"
	yeah *the*.. I thought so too...	'EP'	needed. Although M. cites in support
	that's right... here it says the	(refers	an earlier noun phrase, my
	olympic athletic festival.. that's	to	impression at the time was that this
	right.. let's put *the*.. that's good	title)	was the product of a search (back to
	isn't it?.. THIS ONE IS RIGHT I THINK I		the title) carried out only after A. had
	DON'T KNOW WHICH THOUGH..	Col?	supplied the filler.
	document..I think official		
	document goes back to 776BC		*The conversation shows that (7) is still undecided at this point, at least in the mind of M.—who appears to be contributing most to this recovery.*
9-	A. CHECK?	Log	In the post-task interview, A. could
12	M.	Gra	not recall which point(s) she was
	IT WOULD BE IMPOSSIBLE [IT IT WERE	Tr	suggesting they check.
	NOT HELD ON THE FIELD OR	KOW	*M. has trouble making sense of the*
	PLAIN]...what? but why is it 'was'	Tr	*plural > singular switch. A. supplies*
	here?.. but.. married woman	Ah!	*the correct filler 'no' for (9) in a*
	was... admitted... spectator		*tone that suggests she had already*
	these are all singular what is		*reached the answer without much*
	this?.. THIS I DON'T UNDERSTAND..		*difficulty. M. paraphrases the*
	why is it singular here?.. many		*sentence in the L1 before agreeing*
	thousand spectators..... came		*that A's suggestion makes sense.*
	from Greece but... it requires		
	historical knowledge		Asked in the post-task interview how
	A.		she had arrived at the filler for (9) A.
	ISN'T IT BECAUSE IT'S *no*?		replied that it was "not very

M.
no one... no..ah I see I see..
A.
no married.. then here as *a* ...
M.
as a.. yeah right! THEN IT'S COMPREHENSIBLE.. so married women are not allowed
A:
RIGHT
M.
up to here it's okay..

difficult" Given M's swift recovery of (13), and his claimed prior knowledge of women's restricted status at this period, it is likely that his difficulties with (11) and (12) derived from the plural > singular switch mentioned above. M. had no recollection of having learned this TL construction in school, even though, as A, noted, it is part of the standard English syllabus for Japanese secondary schools.

The extracts above offer some idea of the transparency of most pair-condition think aloud, at least as compared to solo-condition reporting. This must be its main attraction for the researcher, although allowing informants to report in pairs does appear to make it easier (Haastrup 1991; Aizawa, pers.comm.) to attract volunteer informants. My own pair informants, moreover, have consistently indicated (on post-task, anonymous Likert-scale response slips) higher levels of satisfaction with the task itself and with their performance on it than their solo-condition counterparts. T-tests performed on these ratings indicated significant differences in means across the two conditions for GL1 informants (t-value = 2.6201 at df22) and for their JL1 counterparts (2.2193 at df22.) It may also be worth noting here that, to date, among my informants only pair-condition informants have volunteered for further think aloud sessions.

Against the use of pair condition reporting must be set the often considerably greater length of protocols (and, perhaps problematically, the better the paired informants get on with one another, the longer the session is likely to be) and the fact that—apart from PCTA-specific codings such as CP ('consult partner'—the range of coding behaviours and their frequencies in pair-condition protocols do

not markedly differ from those found in solo-condition. The constraints of pair-condition reporting must also have some effect on how the task is carried out, however. It appeared from Anneke and Fred's protocol data, above, that Fred might well have completed the task in less time had he been working alone. He agreed with this suggestion in their post-task interview, but added that (as the final turn, above, shows) he probably would not have filled in so many blanks and would certainly have "said less". Anneke had wanted to spend more time on some deletions such as (7), but felt she had no choice but to agree to move on. She also noted that Fred's ideas were very helpful, so that in the end she might fill more blanks by cooperating closely with him.

In short, one has to balance the pros and cons of solo- versus pair-reporting: solo-reporting may be preferred so as to maximise the number of informants (but see below) yet these may in the end produce less, and less informative, data than might have been gained from paired informants. Last but not least, the already-mentioned low-verbaliser informants may become considerably more productive in paired reporting. I have to date only persuaded two of these individuals to work together in a pair, with the result that they reported enough to allow coding of just over half of the (VIDEO RECORDER) passage deletions. Alone, each had managed to report his processing of less than 10% of deletions on the OLYMPICS passage. Pair-condition reporting, then, may bear serious consideration.

6.6: Pair-informants counted as individuals

Examination of the pair-condition data above reveals that the partners occasionally disagreed about an interpretation or choice of filler. Furthermore, there was not always a balance of input between the partners in terms of successful interpretation and recovery. Observation of pairs at work, however, suggested that apparent disagreements and/or unbalanced contributions may have in part have reflected personal and situational factors distinct from cloze proficiency and/or think aloud ability: gender; mutual perceptions of L2 proficiency; and the Japanese *senpai-kohai* (mentor/seniority) relationship may all have played a role here. My argument, however, is that where informants disagreed about a filler or an interpretation of passage content, this was reflected in their protocol (and on five occasions among Gl1 paired informants by the noting down of two alternative fillers.) In such instances we might reasonably treat the pair as consisting of two minds at work. It seems unreasonable to do so only in cases of disagreement, for consensus about passage meaning or the appropriate filler also requires two minds. I have thus opted to treat all paired-protocols as two individual protocols, noting discrepant fillers, etc. where appropriate.

6.7 GL1 & JL1 Think aloud data compared

I conclude this discussion of think aloud data (though see below for some comparisons between think aloud and the NCR variant of verbal report) with tabulated comparisons of GL1 and JL1 informant data. A table of frequencies with which individual codings were applied to items in all GL1 and JL1 protocols

is shown, for reasons of space and formatting, at the close of this chapter.

Cloze success

The table below compares the mean percentage SEMAC scores of informants from the two language backgrounds on the 32-item (i.e. common to all) OLYMPICS passage, and also breaks the total down into scores on lexical/content deletions and syntactic/structural/functional deletions.

GL1 overall SEMAC	GL1 structural items	GL1 lexical items
72%	76%	71%
JL1 overall SEMAC	JL1 structural items	JL1 lexical items
67%	72%	61%

Figure 6.1: GL1 and JL1 cloze success overall and by two item-categories

A t-test (SSP v2.0) applied to the SEMAC scores of GL1 and JL1 think aloud informants produced t-values of 0.3889 at df26 for lexical items and 0.9791 at df34 for structural items indicating no significant differences in the means scores of Gl1 and JL1 informants. These scores are interesting in that given the presence of seven deleted articles among the structure items (or 39%) in the OLYMPICS passage it might have been expected that–as indeed (cf. chapter 4) JL1 consultants had predicted–JL1 informants would do appreciably worse than their GL1 counterparts (whose L1 possesses an article system) on these items. The fairly frequent comments head in JL1 think aloud protocols about difficulty with articles (cf. Yasuko, above) and/or remarks indicating low confidence in a recovered article filler may reflect a perceived rather than a substantive lack of firm knowledge in this area.

Codable recoveries

The percentage of think aloud protocols which did not contain enough information to permit coding of how the informants recovered a given blank was calculated. A significant difference was identified between informants from the two language backgrounds, with GL1 informants failing to provide insight into their recovery of 7% of deletions (SD 4.9) while their JL1 counterparts did so in 17% (rounded up) on average (SD 11.1). A t-test (SSP v2.0) applied to percentage counts provided a t-value of 4.9391 at df 62, significant at 0.01. The figures thus confirm my own impressions that JL1 informants' provided insight into their recoveries less consistently than GL1 informants, and that the former exhibited a considerably wider range of productivity. One caveat must be mentioned in relation to these findings: my command of Japanese is considerably weaker than of German and, although I consulted a native speaker of Japanese whenever I felt any doubt about the import of what an informant was saying in her L1, there may have been instances in which I failed to recognise pertinent information in an informant's think aloud (or in her retrospective data, which was taken down in note form and thus less open to later checking with native speakers.) This may be unavoidable when informant's report in a language with which the researcher's proficiency is limited, but I would argue that any such omission on my part has biased the content of think aloud far less than would have been the case had I required, as some researchers have done, that informants report in an L2. (See Cohen 1998 for a quite detailed discussion of the choice of language of reporting and the implications of same.)

Common aspects in the think alouds of GL1 and JL1 informants

Even though frequencies varied, for each item, the same dominant or most salient coding was identified in the processing of GL1 and JL1 groups. This suggests that, even where more than one route to recovery exists, there is typically a most likely way of approaching the item, and that this route is available to test-takers of different language backgrounds. Given that so many blanks in natural or fixed-ratio cloze can be filled by reference to learned syntactic knowledge, or knowledge of phrases and collocations, this similarity in processing routes may not be very surprising. Differences become more apparent when the second most common processing events are compared, for here (if 'fallback' operations such as citing the word class of the item are omitted) only some 22% are common to GL1 and JL1 informants. This appears to tie in with the observation above: there may be one most likely route to filling the blank, but a variety of alternative choices can exist. The degree to which a data-elicitation procedure makes apparent these alternative routes or modes of item processing may be seen as a measure of its heuristic value, alongside the above-mentioned comprehensiveness or reliability with which it gathers data.

Extratextual knowledge

The small survey reported in chapter 4, which found a significant differences in German and Japanese L1 respondents' expectations about the role of prior knowledge in a cloze task calls for a comparison of its actual use as evidenced in think aloud data. In the two groups' processing of the OLYMPICS passage, however, there do not appear to be marked differences in the unambiguous

application of extratextual knowledge. Only in deletions (1); (21) and (29) was 'KOW' the dominant coding, and this was the case in both GL1 and JL1 think aloud protocols. A chi2 test (SSP v.2.0) applied to raw frequencies gave a value of 0.6276 at df2, indicating no significant difference between GL1 and Jl1 informants with regard to this application of prior knowledge, reinforcing the initial impression that the two groups did not differ much in their use of prior knowledge on the task.

It seems to be possible to identify three roles for extratextual information in think aloud. Firstly, such knowledge appears to have played a fairly unambiguous role in almost all think aloud informants' recoveries of the three items above, in that without the requisite knowledge it is next to impossible to arrive at the original item, or at a 'narrow' SEMAC alternative. A second and much less consistent role was identified in processing of items such as (11) and (13). Here extratextual knowledge appears have played more of a supplementary or confirmatory role such that the informant was able to use her grammatical knowledge to fill the blanks, but her prior knowledge of the topic may have made this easier or faster, and/or boosted her confidence in her recovery. This may be seen in an extract from GL1 informant Manfred:

"[...] so ELEVEN..mhhmm.. no married woman was admitted even as.. a spectator.. YEAH I KNEW THIS IN FACT.. WOMEN WERE NOT EVEN ALLOWED TO WATCH.."

Given that extratextual information was mentioned, and that Manfred had offered no other indication of how he had arrived at these fillers, this could only be coded

as 'KOW'. Post-task, Manfred was not at all sure about how much of a role his knowledge has played, but could confirm that it had been in his mind "as soon as I [connected] spectators [presumably from the previous sentence] and married women." In other words, this does not appear to have been a straightforward sequence of events involving recovery followed by confirmation from prior knowledge: prior knowledge played some role, but this cannot readily be quantified even by concurrent reporting plus post-task retrospection and is thus probably beyond precise description.

The third role for extratextual knowledge may be seen in informants' processing of deletions such as (7); (18) and (25). Prior knowledge was applied by a few informants only, and in unpredictable or 'objectively' inappropriate ways. GL1 informant Anna, for example, claimed to know (and she may have been entirely correct in this; I do not know) that in Olympia marble tablets (*Tafeln*) were set up to commemorate the winners of certain events. A few other GL1 and JL1 informants hinted at this (cf. NCR Ryou, below) but none in such detail. While neither Anna nor her partner was able to come up with a suitable TL equivalent, she had clearly grasped the meaning of a plausible filler. Two JL1 informants, however, applied extratextual knowledge to some degree in entering the name of a sporting event in deletion (18), whose original item had been 'also'. In taking this route they overlooked the punctuation mark (a comma) that might have steered them away from adding an item to the list of events. More than one informant also queried item (25) 'the fourth day, ... of the full moon' on the grounds that

she 'knew' the moon appears at night. One interesting feature of the NCR reporting format that perhaps merits elaboration here is that all mentions of prior knowledge noted in NCR protocols belonged to the first and second of these three categories, with no clearly 'objectively inappropriate' applications of extratextual information found. I have no real idea why this was the case, but I would speculate that only salient or 'valuable' extratextual cues survive (see below) in NCR informants' reports, or inferences. One interesting and slightly ambiguous case can be found in Ryou's processing of item (27). Here he explicitly mentioned at least the suspicion that the ancient Games lasted six days, but in the end chose not to apply the information on the apparent grounds that cotextual information was adequate. My interpretation of remarks by 'focus group' AC informants (cf. chapter 7) is that, where the two types are available, textual cues may take precedence over extratextual and this fits with Ryou's post-task comment that 'the fifth and last day' seemed a safer option in item (27).

Translation into the L1

If the role of extratextual knowledge in cloze item recovery is not always clear, this is even more true of translation behaviour. Although a good deal of L1 glossing goes on in the verbal reports of think aloud informants, sometimes clearly audible but very often in an apparently self-directed 'mumble', it is seldom easy to know how much of a role this is playing in recovery as such, and how far it represents an economical structuring or 'gisting' of the overall passage content for easier recall or access to meaning. At times, however, translation is carried out

at length and with care, and in these cases it seems to be playing an unambiguous role in recovery. My hypothesis had been that translation was *typically* an indication of difficulty in filling a blank, or in comprehending a span of passage content (but cf. my discussion in chapter 7 of graphic marking on AC manuscripts of passage content translated) althought this difficulty need not be reflected in the informant's cloze success on the item(s) concerned. Where translation succeeds in telling the informant what she wants to know, cloze success is of course quite likely. Overall, the two items in OLYMPICS which stimulated the most extensive and sustained translation behaviour by both GL1 and JL1 informants were (14) '..the exact ..... of events..', and (25) '..... of the full moon.' Both of these seem to have posed a fairly high level of challenge to many informants, but protocol content and post-task interview comments suggest that the reasons were different in each case. Although many informants used phrasal or collocational knowledge (cf. chapter 7 and AC data) to come up with the SEMAC filler 'exact number', some of these continued to translate the span, as in GL1 informant Alex's extract below:

"WELL.. exact SOMETHING of events.. SO SOMETHING IS BEING DESCRIBED EXACTLY PERHAPS... OR IN DETAIL.. the exact number of events MAYBE? [IMPLIES HE KNOWS THIS PAIRING] the exact kinds of events YOU DON'T SAY IN ENGLISH I THINK.. the exact list?.. COULD BE BUT.. UNCERTAIN IS the exact [INAUDIBLE] AHA.. the exact schedule PERHAPS? TIME-WISE EVERYTHING HAD TO BE PLANNED VERY EXACTLY.. the exact schedule I'D SAY.."

Post-task Alex confirmed that he had indeed translated with a view to recovery, and claimed to have been aware "more or less from the beginning" that 'number' might be acceptable but that he had been looking for a better alternative (which in

'exact schedule' he felt he had found.) A similar comment was made by a JL1 informant, who also claimed to have thought of exact number' but also suspected that this might be "too easy" a choice.

Translation seems to have played a more primary and direct role in the case of the other item mentioned, (25). This blank clearly posed created serious difficulty for a good number of informants, and in several protocols the span translated extends from deletions (22) to (26). Post-task, some of these informants confirmed that they had translated extensively in an effort to understand the sentence structure of the passage context, and one GL1 informant noted that although ideas for possible fillers had come to mind, he did not feel able to choose among them until he had grasped the grammatical sense. I will return briefly to the topic of translation behaviour in cloze in chapter 7, where I discuss the use of graphic devices by means of which AC informants can indicate with some precision those parts of the passage which they translated. I would note here, however, that although some of the translation that goes on during thinking aloud appears to be 'suppressed' (i.e. taking the external form of inaudible or uninterpretable 'mumbling' which we must suppose is highly self-directed) the data I have gathered suggests that translation has a variety of uses. These include getting to a candidate filler, choosing among alternatives, and (apparently only in a very few cases) even evaluating how well fillers go together in a wider sense.

6.8 Non-continuous reporting (NCR) defined

In chapter 7 I discuss the use of a variant non-verbal reporting procedure I have labeled 'annotated cloze' or AC. My attempt to develop an alternative to think aloud stemmed from the realisation that this procedure was fairly unpredictable, highly labour intensive for the researcher, and often fatiguing and frustrating for informants. My explorations of AC revealed problems as well as benefits, however, and it occurred to me that there might be some way of retaining the advantages of think aloud while downplaying its negative aspects (cf. Cohen 1998:36ff.) Cognizant of the apparently slightly greater leeway in Ericsson & Simon 1993 in terms of how immediate a verbal report must be in order to be seen as reliable, I opted to investigate the use of a task format in which informants are asked to report how they recovered a deletion just as soon as they have done so. If the interval between the processing event and the report is no more than ca. 20 seconds, NCR may be taken to represent what Cohen 1998:49ff labels 'self-revelation.' Where the interval is longer, it becomes more like what Cohen (ibid.) calls 'self-observation.' A simpler conception of NCR is to see it as a form of retrospection with variable delay. For some cloze items, NCR reports occur almost at the speed of think aloud—which is to day that recoveries may appear instantaneous—while for an unusually challenging blank a minute or even more may elapse between 'encounter' of the blank and commitment to a suitable filler. This range echoes that observed in think-aloud.

Process vs. product?

Although at times superficially quite similar to think aloud, the differences between it and NCR should not be underestimated. The essential requirement of the former can be summed up as "Tell me what you're doing", and of the latter as "Tell me what you just did." Think aloud, moreover, tracks the process of looking for the answer, while NCR (and, arguably, any other form of retrospection) looks more at how it was found. This distinction may not be trivial. Taft 1991 notes as uncontroversial the notion that once a solution is found to a problem, the details of the 'cognitive route' to that solution may be rapidly forgotten. It is plausible that something along the same lines will occur when informants are asked to report how they arrived at their chosen filler for a cloze blank. NCR reporting, then, might 'lose' data that think aloud would be more likely to retain.

But would this be a good or a bad thing? Again, the question is not a trivial one. I have mentioned in earlier chapters the 'self-directedness' of much think aloud data, and the difficulties this can bring to the interpreter. It would be no bad thing if what got reported were in a more readily understandable form, and it was my anticipation that by taking away the expectation that an informant speak more or less continuously it might become easier for her to order her thoughts, to edit out those steps or events less salient to recovery of the filler, and to structure what she had to tell in a more 'other-directed' way. Some information might indeed be lost, but–as in a précis of a text–what remained might be sufficient as well as more accessible. A 'tighter' verbal reporting format might allow the collection of data

from larger numbers of informants, moreover, and it might also be more amenable to those (the 'low verbalisers' already mentioned, and taken up again in the discussion of AC procedure in the next chapter) apparently not cut out for the conventional think aloud task.

6.9 Outline of NCR reporting procedure

10 informants for the NCR data-elicitation were recruited by direct invitation from among my JL1 undergraduate classroom students rather than by a snowball sample–although as two or three individuals took some time to make up their minds about whether or not to take part the decisions of acquaintances cannot be ruled out as a factor in recruitment. Six of the informants were members of a test-preparation group, and so their scores on practice tests could be compared. These were found to be quite closely matched in terms of reading sub-test scores, and broadly similar overall; some variance was noted in impressionistic assessments of oral production, however.

NCR data-elicitation session were scheduled within three days at the beginning of a vacation period–in part because informants were less likely to encounter one another on campus at this time and thus less likely to discuss their experience of reporting. The other step taken to prevent leakage of information about the NCR task was to ask informants if there were any topics they especially did not want to read and report about, and by this means to create the inaccurate impression that individuals would be matched to an appropriate passage topic chosen from a set. I subsequently noted no indication that confidentiality had been compromised in

any way. A number of NCR informants asked for reassurance that their performance on this task would not affect their semester grades, and this was given. They were told that, if they wished, they could find out their own scores on the test as well as the average score for the group test. Also, if they wished, I would go over their task sheet with them afterwards and explain why a given answer was right or wrong, and try to answer any questions they had. Four informants took me up on this offer, and these 'de-briefings' were carried out the following week. (Anyone considering the collection of task-based data from learners might note that three of these four indicated that the offer of a 'learning experience'—which I accept had been tagged onto the task as a form of bait—had been their main motive for taking part.) The same principles of confidentiality applied to think aloud data (recordings were transcribed and then erased; task sheets were identified only by number, etc.) were applied to NCR.

The pace of technological change in Japan is such it is no longer safe to assume that everyone can operate a cassette recorder. With this in mind, and anticipating that NCR reports would take the form of discrete spans of verbalisation interspersed with silences, I opted to provide my informants with a digital MD recorder which allowed the user to create a new track simply be pressing a button. Informants were instructed to do so immediately prior to each report stage. Their reports of how they recovered deletion (1) thus became track 1, their reports of deletion (2) became track 2, and so on. This feature makes for much faster auditing and the easier location of items of interest. The informant may also

quickly and easily review her reports, and add to them if she so wishes and the researcher allows. (Several digital recorders allow the insertion of extra information into a recording without loss of existing data.) It was of course possible that informants might verbalise 'informally' between the required reports, and so in the earlier sessions I also arranged (see below) for a continuous recording to be made. No special training or model was given to informants, who were simply instructed to:

> *"Try to complete the cloze passage on this page by writing* ***one*** *English word in* ***every*** *blank space. As soon as you have chosen a word, press the red button on the recorder and say as much as you can about how or why you chose that word. Please also mention any other things you noticed or thought about as you were choosing the word. If you thought of more than one word, please mention them all if you can. You may speak in Japanese, in English, or in both languages. When you have finished speaking, press the button again. Repeat these steps every time you wish to report about how you chose a word for a blank space."*

Although this instruction was prepared only in English (with graphic illustrations) informants' grasp of the task was checked in their L1. None expressed any uncertainty about what she had to do. As noted above, I anticipated that, in the absence of any task-requirement to keep up a steady flow of verbal output, what NCR informants did report would be in a better-structured and more readily interpretable form. It would, I expected, also be more condensed and thus require less time to transcribe. But I also anticipated drawbacks, the chief of which was that worthwhile information might be verbalised between reports, as it were, and thus be lost. For that reason, the accompanying continuous recording was not under the informant's control. The fact that a second recording was being made

was not hidden from informants, but the directional microphone was set out of their line of sight lest they forget to use the digital recorder to hand.

I observed all NCR sessions, and found that without any need to watch the clock for extended silences, as in think aloud, it was easier to make notes about NCR informants' behaviour. (As the NCR recording format does not automatically track the informant's time-on-task, however, some means of noting starting and finishing times must be used: a simple digital stopwatch is adequate, and may be operated by the informant herself if unsupervised.) In the event, the backup continuous tape recording proved to be of little value. In only a few sessions was there anything of interest on the tape at all (i.e. beyond the verbalisations already present on the informants' own recordings) and virtually all of this I had been able to note down. Some of this content was also reflected in the NCR reports ("I chose the one that sounded best"; see below) but some was omitted. In the ten NCR sessions, I noted a total of 11 instances (five of which occurred in a single informant's session) of what I interpreted as the sounding out, between reports, of candidate fillers (e.g. "the plain of the plain near the plain at mount Olympus..") as well as several L1 exclamations of difficulty ("DIFFICULT!") and sudden reinterpretation ("AH, I GET IT!") Although these did not appear in the informant-controlled NCR audio-protocols, I did not feel that they would have justified the auditing of some three-and-a-half hours of continuous audio-recording. In other circumstances recording in full may be desirable or indeed the only choice. In a group reporting session (and there is no reason why

NCR should not be conducted in a regular LL) a continuous recording would track any verbalisations informants made between their self-initiated reports–something no single auditor could reliably do.

6.10 An example NCR protocol (Ryou)

It may be as well at this point to offer an illustration of what an NCR protocol looks like. Below is the virtually unedited NCR transcript of one male Japanese L1 informant's NCR session.

JL1 NCR informant Ryou

Ryou had spent three or four years in the USA as a boy, returning to attend a Japanese high school and university. Ryou's oral TL proficiency was fairly high, and he chose to report in English. His few L1 interjections are mainly L1 equivalents for the chosen TL fillers, or glosses of extant passage content. The references to numbers "..one maybe.." represent Ryou's ratings of individual item difficulty on a 1—5 scale (see below) with which he was familiar from classroom activities. (Only some 40% of items were rated here.) Although Ryou completed the full 41 deletion passage in less time than regular think aloud informants took for the 32-deletion version, I have followed my practice above, and show below only the first 32 items' worth of protocol data.

Del.	NCR protocol data	Codings	Post-task comments
1	well.. it's every four years.. I know the games happen every four years now and I think it was the same those days. yeah..one maybe...	KOW	
2	okay.. I think this is something like changed their [sic] local character.. CHARACTER IS (*SEIKAKU*) RIGHT? I can't think of another way to say in one word..	??? L1Eq	Ryou could not elaborate on his choice here, which he felt sounded "strange". He agreed that he had glossed the span in his L1
3	national can go here. It was local then national and then later on it became international.. so national makes sense in between them.. national AFTER THAT international..so two maybe..	SSb/ Log SSf	
4	against.. rules against foreign competitors.. first there were rules against foreigners and then they changed them and.. then foreign sportsmen could enter the games..ah..one	Phr/ Col? TLPara ph	Ryou confirmed that he had learned the pairing 'rule against'
5	no one knows.. I thought that it had a.. a hyphen but maybe not.. anyway it's a common word [sic]	Phr LAN	
6	it must be the.. you need an article here..I thought of ancient but then you still need to put an article SO..one	Gra	
7	I thought of documents but I don't think they had paper in those days.. in Greece I mean.. so I put official records.. or it could be like.. INSCRIPTIONS? In.. in stone.. records is okay I think	KOW Log?	Ryou noted that 'records' could include paper documents, inscriptions, etc.
8	took place.. I know this kind of verb with a.. a preposition.. you learn them in school	Gra KS	
9	well.. it has to be a preposition so maybe on or near . . ah.. I'll choose near because you don't find a plain on a mountain.. that's STUPID	WC CBA Log/ KOW	
10	came from all parts of Greece is okay here.. you could use another verb like traveled but came is good.. ehh.. two?	???	Ryou "just kn[e]w" the item 'come from'
11-12.	I put no married woman was allowed even as a spectator.. I learned this kind of sentence.. CAN'T REMEMBER WHAT THEY'RE CALLED.. and I knew that married women couldn't watch the games.. BECAUSE ALL THE ATHLETES WERE NAKED OR SOMETHING..	Gra KOW	
13.	were not allowed to compete.. you need be and it's persons so were not allowed.. THAT'S ONE FOR SURE..	Gra	

14.	okay this could be like schedule or list.. or number could go in.. exact number sounds good so that's what I put in..it's two or three..	CBA Sou	Ryou felt that 'exact number' sounded more natural than 'exact schedule', etc.
15.	well.. it's like a list of sports here so I think included.. I can't think of another verb to go here.. FOUR I GUESS..	SSf Log	
16.	FOR EXAMPLE in Japanese.. such as.. when you are giving examples..	???	Ryou "just kn[e]w" the item 'such as'
17.	I think some is okay here.. it says foot race*s* so some..	Gra	
18.	I think even goes in here.. I didn't know they had boxing.. I thought the Olympics was all about peace.. and they stopped all the wars.. so there was even boxing tatata..	KOW	Ryou indicated that 'even' matched his surprise that boxing was part of the games.
19.	I put of here but it looks funny.. of varied ability.. don't you say of various abilities?.. VARIOUS ABILITIES RIGHT? WELL I ONLY HAVE TO PUT A WORD IN .. so [I put] of varied ability	LK LAN L1Eq	Ryou claimed to have chosen of in the expectation that the same preposition would occur with 'varied' and 'various'. (See below)
20.	okay this must be the..the winner.. I know that in the pentathlon there's only one winner.. whoever gets the most points in all the events or something..three..	KOW	
21.	Okay.. he was the champion in running so this must be another sport.. and I'm almost sure it's a kind of jumping.. I don't know if it's high-jump or long-jump so I put jumping.. maybe six? THAT'S A JOKE!	Log KOW	Ryou was unsure whether 'high-jump' was written as two words, with a hyphen, or as a single word.
22.	this must be the because you need an article here too..one..	Gra	
23.	well this must be a verb.. I know the meaning like in Japanese..DAY FOR SOMETHING.. but I can't think of a good English word.. I'll come back to it..	WC L1Eq LFN	In the PTI Ryou spontaneously came up with 'dedicated'.
24.	I know heroes of the day.. like soup of the day [LAUGHS]	Phr/ Col	
25.	The only thing I can think of here is day again.. it looks strange.. the fourth day.. day of the full moon.. five anyway.. maybe I'll come back to it..	LK LFN?	The gist of Ryou's comments appeared to reflect discomfort at reusing a word which had just appeared in the passage.
26.	I'm not sure about this one.. I think set aside can go here.. set aside as a holiday.. eh a holy day.. like it said about religious.. religious associations.. so set aside as a religious day..	Phr/ Col EP	Ryou confirmed that he was referring to the mention of 'religious associations' in the first sentence.

27.	okay.. it said fourth day just before so I guess the next day is the fifth day.. or maybe the Olympics is six days? It must be like a number.. fourth fifth sixth..too many numbers..	SPb Log KOW?	Ryou claimed to have considered entering 'sixth' but thought 'fifth' was "safe". (See below)
28.	All the victors were crowned.. you need be here..so.. one?	Gra	
29.	okay they had like.. the winners got these crowns of something [GESTURES TO HEAD] THE JAPANESE WORD IS *GEKKEIKAN* (laurel crown) .. or is it olives.. like olive leaves? I WONDER (*SAA*).. maybe I'll come back to this one..	L1Eq KOW CBA LFN	Ryou indicated that he had thought of 'olive crowns' but was distracted by the better-known L1 equivalent of 'laurel crown'. (See below)
30.	okay this is like.. a backwards.. so SOMETHING was the SOMETHING that dadada... it should be so great was the honour.. I learned about it..umm.. two I guess..	RFL1 Gra KS	
31.	the could go in here but I think each is better because it says there were *some* foot races..	Gra EP	Ryou confirmed he had been referring to the span following (17)
32.	the I think is good here.. or.. can you say that year of his victory? sounds strange so maybe the.. WELL you need an article or.. I DON'T KNOW THE NAME..OH..THREE OR FOUR..	Gra LAN	

Ryou's time-on-task was just over 21 minutes, and his overall SEMAC score was 85% (87.5% on the first 32 deletions.) Ryou asked to go through his reports again, but agreed that we would audit them together so that I could put questions to him. His post-task comments are shown in the right-most column, above.

Discussion of Ryou's protocol

Even though Ryou was a fairly able informant, overall, there were four occasions on which he offered no on-task insights into his recovery. In three of these, however, the post-task interview helped clarify matters. Ryou's post-task comments suggest that some worthwhile insights do indeed 'fall through the cracks' between reports, so that several of his post-task observations are worth looking at in detail. We could perhaps have inferred from the NCR content for deletion (7) that Ryou was using 'records' as a superordinate term, but this was

made explicit in the post-task interview. The protocol content for deletion (18) also allows us to infer Ryou's reasons for choosing 'even' as the filler, and his post-task remarks merely reinforce this. His comments about deletion (19) are perhaps more valuable in that he makes clear his expectation about prepositional usage. His comments about deletions (21) and (27) indicate either a degree of caution, test-wiseness, or both. Ryou's post-task elaboration regarding deletion (25) suggests that a factor in the 'strangeness' of "day of the full moon" may be the notion, seen also in Yasuko's protocol, that there is something not-quite-right about repetition of TL words in close proximity. As noted above, Ryou's sporadic mention of numbers was occasioned by my request that he rank each recovery on a difficulty scale of one to five. (This scale was not physically present on Ryou's task sheet, which may have been a factor in his ranking of only ca. 40% of items on the scale. The six informants whose sheets included the scale performed better, at ca. 65% on average.) There does not appear to be a pattern to this aspect of the data, as Ryou sometimes rates and sometimes omits both (apparently) easier items and (apparently) more difficult ones.

It was useful to have confirmed, post-task, Ryou's apparent references to passage content at some distance from the target deletion, as in (26) and (31). My observation notes made during the NCR session reveal that Ryou appeared to scan backwards in the passage (although he was unable to recall this post-task) at least three times in the latter half of his processing, and on each occasion just prior to making a report. I also followed up his post-task remark that he thought he had

processed together deletions (38) and (39) by asking him to mention any other items he felt he had filled in tandem. Similar item-linking was noted in the think aloud protocols of Claudia and Yasuko, and other informants, and I wondered if its apparent near-absence here might be an artifact of the non-continuous reporting format. In the event, Ryou did not think he had 'co-processed' any other items, but in any future rounds of NCR data-elicitation I would try to re-phrase the task instructions in such a way as to legitimise 'chunking' of items and reports. There is, of course, the risk that doing so might cause more information to get lost between more widely-spaced report stages.

Another interesting item gleaned from the PTI was Ryou's explanation of why he had rejected 'olive(s)' in deletion (29). He clearly had an inkling that this was a plausible filler, but the strength, as it were, of the visual and/or orthographic image of '*gekkeikan*' or laurel crown led him not to enter his alternative. The NCR protocol makes clear that he was weighing two candidates, but does not tell us why the deletion was (temporarily, at least) left unfilled. For what it is worth, the notion that Olympic winners were crowned with laurel wreaths appears to be so plausible to English native-speakers that to date none of those whom I have consulted has questioned 'laurel' as a SEMAC filler for deletion (29) in the OLYMPICS passage.

6.11 Post-task interviews in NCR

Given that one of my objectives has been to find more efficient ways of gathering data about cloze processing behaviours, I felt a certain ambivalence about the

insights, above, from the post-task interview. A post-task interview is often indispensable when think aloud data has been collected, but my hope had been that it might be superfluous in NCR. This hope derived not only from a wish to save time, for by no means all informants appear to welcome or enjoy the task of reviewing their audio-protocols–many of us, after all, cringe at hearing ourselves on tape–and answering questions. Some may claim to be, or genuinely be, unable to recall much about their task behaviour at all, and at its worst the post-task interview can be a sterile or even prickly formality.

That said, my experience of post-NCR interviews has been comparatively positive. First of all (and cf. section 6.14, below) the fact that NCR verbal report data is less fragmented and in large part 'other-directed' means that there is less that needs to be clarified, and the informant herself may find it easier to elaborate about clearer data if this *is* required. Interviews thus tend to be briefer and, for want of better terms, less convoluted and taxing. The less ambiguous information that NCR protocols contain meant that this post-task interviewer felt able to put questions about finer aspects of processing (cf. 6.14) without much risk of biasing the informant's retrospections – a potential problem if, as more often occurred with thank aloud data, interpretations seemed less well-grounded. The speedier access made possible by digital recording media (cf. 6.9) to the points of interest I had noted also contributed to making NCR post-task 'interviews' considerably shorter than those following think aloud.

6.12 Some pros and cons of NCR procedure

The NCR variant of verbal reporting does indeed appear to have several advantages over think aloud in terms of 'efficiency'. The time required to complete the task was considerably lower for NCR informants (using the 32-deletion task as a yardstick: the point at which all NCR informants filled item (32) was noted) but this is hardly surprising or meaningful given the divergent task requirements. As anticipated, the time required to transcribe NCR protocols has been in my experience markedly (ca. 40%) lower than that needed for think aloud data, and far fewer queries regarding informants' L1 verbalisations have had to be taken to NS consultants. (Note that this is not simply a function of language of reporting: NCR informants' L1 verbalisations also appear to be reliably more interpretable than was the case in think aloud.) As mentioned above, post-task interviewing is a less demanding process: informants appear less fatigued, and appear able to confirm or comment on their behaviours more readily on auditing their own protocols. Also as anticipated, however, it appears that certain events in the processing of cloze found in think aloud protocols are not recorded by their NCR counterparts. Examples are cited in section 6.13, below, but it may be felt that the non-appearance of such extremely common behaviours as 'running fillers' (TL or L1 something, sounds) is not a serious drawback: we may infer from protocol data that the great majority of cloze test-takers use these running fillers at least some of the time–although I accept that for purposes such as quantifying L1 vs. TL use in reporting it might be useful to know whether such fillers had been in the TL or the L1.)

Another item which NCR may record less effectively than think aloud is the difficulty level the informant experienced in filling a given blank. Some NCR informants (cf. Ryou's protocol, above) did verbalise about how difficult a particular item was, but this is not reflected as consistently as in think aloud data presented earlier. Fortunately, the lower (or less sustained) cognitive demands of NCR seems to provide greater leeway for the inclusion of additional task dimensions, such as (as used in AC; see chapter 7) the addition of a five or six value scale of difficulty which informants may be asked to complete for each item. Through an oversight, only six of the ten NCR informants were set this additional task, but among these overall some 65% of the item difficulty responses were completed. It is difficult to compare in a useful way these selected scale values with the far less precise indications of difficulty found in the think aloud protocols of Yasuko and others, above. It should also be borne in mind that only think aloud can provide, as it were, 'snapshots' of perceived difficulty *during* processing. NCR informants offer comment or base their scale selections on a final perception of how difficult it was to recover an item, and were a think aloud informant to reinterpret passage content in such a way that recovery suddenly became easier, her final impression of the item's difficulty might be skewed by that event. Absent very close observation, only her audio-protocol data could provide a corrective.

6.13 JL1 think aloud and NCR data compared

To date I have been able to gather only 10 NCR protocols based on the OLYMPICS close passage (all from JL1 informants) so that inferences drawn from this data can only be very tentative. Moreover, a chi2 test (SSP v2.0) on the relevant items in informants' pre-course questionnaires revealed that NCR informants were significantly more likely to have spent time outside Japan than their think aloud counterparts. Given that this usually means time spent in an English-speaking culture or school, it is not altogether surprising that a t-test (SSP v2.0) revealed a significant difference at .05 in the means of the two groups SEMAC scores (t-value 3.8435 at df20.) In other respects, however, NCR informants could be compared to their think aloud counterparts. Despite their better TL proficiency, four of the ten NCR informants chose to report mainly in their L1, and (although NCR informants' time-on-task was inevitably lower) my impression was that their processing times for 'easier' and 'more challenging' items corresponded roughly to those of JL1 think aloud informants.

What can realistically be compared?

Some anticipated divergences in the kinds of data recorded by think aloud and NCR were discussed above, but which of these might merit closer examination? Because I had transcribed the protocols separately as I obtained them, I had no more than loose impressions of the relative frequencies with which individual processing events occurred in NCR data, and before making any attempt to calculate these I tried to come up with a set of processing events for which I could see no plausible explanation of frequency differences in terms of task factors.

That is, I looked for behaviours that I intuited should occur roughly as often in one task condition as in the other.

It was easier to construct such a list by leaving out those behaviours and events whose anticipated lower frequencies in NCR protocols could plausibly be explained as artifacts of the procedure. The already-mentioned lower incidence of 'item grouping' in NCR is one instance, for in NCR reports recoveries are in principle isolated events. 'Running fillers' are another type of processing event that we would not expect to find in NCR protocols, as these appear to be a standard 'mechanical' device in maintaining the flow of passage content (cf. Allan 1992a:175) and are perhaps only barely 'deliberate.' Overt 'extensive' translation (a pre-recovery 'making of sense' of troublesome passage content) is something else that we might not expect NCR to record—although L1 glossing at the levels of word and phrase might be present in the form of referenced first language equivalents (L1Eq.) Whether this merits inclusion in a list of events worth comparing across conditions is debatable. On the one hand, L1Eq events often appear incidental to recovery, with no real indication that an L1 glossing step made any real contribution to filling a blank (cf. deletion (2) of Ryou's protocol, above, in which he appears to cite the L1 equivalent in a moment of uncertainty about his choice of filler) while at other times an informant apparently offers an L1 gloss to indicate understanding of the meaning of a filler she cannot supply.

"[...] well.. it's something like order.. the order of events.. I don't know English word [...]" JL1 NCR Natsumi

Incidental glossings appears to be the more frequent, although it is very difficult to be sure of the status of these events, so that I reluctantly exclude these from comparison. Paraphrasing of a span of content in the TL is another event seldom encountered in NCR, and for the same reason: clarification of passage meaning comes before identification of an appropriate filler. (Exceptions might be those occasions when the informant cannot fill a deletion; in such cases she might have as much interest in 'saving face' or demonstrating some understanding of the context as a 'regular' think aloud informant.) The coding 'Und', indicating that the informants claims to understand the meaning of the missing item and/or its context is also left of the list here because in the absence of an appropriate filler it is difficult to be sure that the informant actually has grasped what she claims.

Certain events appear less likely to be affected by task condition. Assuming equal TL proficiency (see above) there was no reason to expect that either think aloud or NCR informants would be more likely to fill OLYMPICS content deletions such as (7) and (14), etc. appropriately. Each group, then, might be expected to have equal recourse to the 'partial-success' marker of identification of a deleted item's word class (WC). The NCR reporting format, however, makes mention of an item's word class relevant only where the blank could not be filled. Moreover, it is not always clear in think aloud whether intuition of the item's word class was a stage in recovery, or merely incidental, and together these factors make it hard to see how the frequency with which think aloud and NCR informants mention word class can meaningfully be compared. The same criterion may be thought to apply

to the behaviour of seeking a single-word alternative (SWA) for a phrase or chunk candidate filler disallowed in cloze. Another candidate for comparison here would be the elaboration of 'knowledge sources' (KS) but while this might be of central interest in a more sociologically or pedagogically-oriented study, it is already fairly clear whence JL1 informants derived the bulk of their knowledge of TL syntax and vocabulary.

Given that candidate fillers appear to present themselves with varying degrees of plausibility in think aloud, and to persist in the informant's 'working memory' (cf. Taft 1991) for varying lengths of time, it might be expected that only the most plausible alternative filler, if one existed, will be mentioned in NCR data. How often this occurs is perhaps less interesting than which items it happens to, and for that reason I do not compare frequencies of 'choosing between alternatives' (CBA). The noting of unfamiliar items of passage content (UPI) is present in think aloud data, but appears to be suppressed in NCR reports: only one of ten NCR informants saw fit to mention the extant passage item 'waived', even though only two of ten claimed post-task to be familiar with it. Other low-frequency items which at least momentarily exercised the minds of think aloud informants, but which were seldom or never mentioned in NCR protocols were: 'discus', 'pentathlon', 'sacrificial', and 'garlands'. Most NCR informants claimed, post-task, to have inferred at least an approximate meaning for some or all of these, and in some cases (as evidenced by their immediate ability to produce an appropriate gloss when asked) to have settled on an L1 equivalent. Asked why

they had not mentioned these unfamiliar items, some informants pointed out that as far as they were aware this had not been part of the task. The instruction had been to "[...] say as much as you can about how or why you chose [the filler] word [...] also mention any other things you noticed or thought about as you were choosing [it]" As with the issue of item-grouping or 'coprocessing' in NCR discussed above, in any future use of the procedure one might usefully amend the instructions to invite mention of any unfamiliar passage items which interfered with the task of filling blanks, or to which the informant can recall devoting attention.

What, then, is left? Events and operations which I assume to be fairly independent of reporting condition and thus potentially comparable include those 'source of information' identifications coded as 'Phr', 'Col', and 'Idi'. These I propose to lump together on the grounds that many informants appear to use a single preferred label—perhaps the only one they have learned—for all of them. (The clarity of these categories is also a matter of debate, cf. Nattinger & DeCarrico 1992.) In recovery terms all three seem, along with 'Gra', to reflect the activation of information from the immediate context of the blank. Activation of grammatical knowledge might be expected to get reported in both conditions, as there are a number of instances where this appears to be the only possible route to an appropriate filler. 'Gra' is not coterminous with 'Phr/Col/Idi' even though both codings represent activation of local context cues. Moreover, although I have found a number of instances in my AC data (see chapter 7) in which informants

coded recoveries like '...to the (24) heroes of the day, ...' as 'Gra', such clearly syntactic items as '... all the victors (28) were crowned with holy garlands..' were not coded as 'Phr/Col/Idi'. Does either think aloud or NCR more comprehensively track activation of these local context cues?

Activation of extratextual knowledge in recovery (KOW) appears from think aloud protocol data to be a cognitively prominent and thus highly 'mentionable' event, and thus open to comparison across the two reporting conditions. The status of 'logical inferences' (Log) is less clear, for as examination of Ryou's protocol will show, there were events which seemed to rely to some extent on extratextual knowledge as a trigger or support for the inference, making it hard to choose the more appropriate coding. For this reason, and because the making of logical inferences is not always identifiable in think aloud data, I propose to omit them from comparison here.

Questioning the need for a filler in a given deletion is a behaviour fairly commonly found in think aloud protocols, and one which suggests a certain sophistication in the informant's understanding of the passage and in particular of valency and/or lexical collocation. Whether or not this interesting behaviour survives in NCR might, then, be a criterion of its value as a data-elicitation tool. How likely is this? From the think aloud protocols of informants such as Claudia, and especially Yasuko, I have the impression that these questionings of the need for a filler were cognitively and/or affectively powerful events, and if this is so then the behaviour should persist even in the edited reporting found in NCR. On

the other hand, under NCR conditions the report follows the choice of a filler; if the informant has already identified an appropriate filler, why should she question the need for one? In fact, as we see from Ryou's data, above, the 'Nec?' event does appear to be recorded by NCR, although it is plausible that in the case of deletion (17) think aloud procedure would have left some concurrent trace of Ryou's doubt, which apparently was not strong enough to merit a mention in his NCR report, as occurred in deletion (37). That said, questioning of the filler in (17) is rarer in both reporting formats, while (37) appears to be more marked in this sense. (I note here, too, that the original filler in (37) was 'training period'. While this collocation persists in current military and bureaucratic English usage, the use of 'training' as a noun may be replacing its use as a modifier in many other spheres, as in 'You undergo three months' training.' 'Worthwhile', of course, has long been used without 'seem' or 'feel' in contexts like: "It's the vacations that make the job worthwhile." A brief survey of textbooks suggests that these verbs do still collocate with 'worthwhile' in Japanese school English, however, so that a much larger study might trace an association between sources of TL knowledge or exposure and perception of 'unnecessary' items in suitable gap-filling tasks.)

This brings us once again to the issue of long-range textual constraint in cloze. As already discussed, the think aloud evidence for uptake of more distant textual cues is slim. Ryou's NCR protocol, above, contains two mentions however (in relations to deletions (25) and (31)) of the use of information from outwith the

paragraph. The codings 'EP' and 'LP', then, denoting the use of content from an earlier and later paragraph, respectively, may be worth including in a comparison of what is mentioned by think aloud and NCR informants. The same is true for codings of information uptake from within the paragraph, i.e. SPf and SPb. As backward reference' within text is thought to be more common than forward reference (Pressley & Afflerbach 1995) these codings' directional distinctions should be preserved. For reasons of space, the table showing frequencies with which codings 'WC', 'SWA', 'Nec?', 'Gra', 'Phr/Col/Idi', 'KOW', 'SP', 'EP', and 'LP' occur in think aloud and NCR is shown at the close of chapter 7 (Figure 7.14; this table also includes data from AC manuscripts.) and according to the percentage by item of think aloud and NCR protocols containing data coded under these headings. Points of note include those discussed below.

Comprehensiveness of think aloud and NCR

Figure 7.14 (in chapter 7) summarises in tabular form information derived from think aloud, NCR and AC informants data, while more detailed comparisons of the codings these reporting formats produced for individual cloze items are presented in Appendix 4. The reliability of the three formats' elicitation of codable data is taken up in section 7.20 of chapter 7, but I will briefly compare think aloud and NCR here. As figure 7.14 reveals, NCR appears to be superior to think aloud in terms of the comprehensiveness (or reliability, with a mean of 15.3% of deletions uncodable from think aloud data but only 5.8% in NCR. To some extent this may be seen as reflecting as much a shortcoming in think aloud

as a strength of NCR, for in the former we find that some items, (often grammatically-cued, and recovered perhaps too speedily and/or with too little processing to be reportable) are filled without leaving an indication in the data of how recovery was achieved. Unsurprisingly, in view of the focus on grammatical rules and form in Japanese school English education, these easier items tend to be auxiliary verbs and in some cases articles.

Although NCR is not productive in every instance, we should not be surprised that it more reliably produces codable data. As I discuss in chapter 7, my speculation is that one key reason for the discrepancy between think aloud and NCR is that in the latter informants are able to (and required to) 'suspend' or interrupt their processing in order to make a report at each recovery; this, post-task discussion has confirmed, is a comparatively difficult point to overlook. The think aloud informant, on the other hand, has the ongoing and arguably more challenging task of reserving part of her attention for the task of reporting her processing operations while simultaneously trying to fill the cloze blanks, and—again, as post-task interviews confirm—informants may find this extremely challenging.

In connection to 'unrevealed' events in think aloud, it is difficult, perhaps impossible, to know whether the ability of some informants to retrospectively (i.e. in post-task interviews) call up events or inferences from their processing of such items represents task-recall or after-the-fact (re)construction. When an informant says something like

however, it does not seem implausible that she is simply noting an item of grammatical knowledge which played a role in her recovery but which somehow did not get reported at the time. When the concurrent data leaves no corroborative trace, however, there is no obvious way to verify an informant's retrospective claim.

Agreement in codings applied

How far we may reasonably compare think aloud and NCR in terms of which codings are drawn from the raw data is open to debate. As discussed in this chapter, there are processing events which are intuitively likely to be better recorded by one procedure than the other, if recorded at all, and these are not realistic objects of comparison. What figure 7.14 (in chapter 7) does reveal is that in just over 90% of possible cases, the same coding was most frequently applied to informants' data; this suggests that both formats are able to track the most salient processing events fairly consistentl. In only 34% of instances were the same *two* event codings most commonly applied, but this must to a large extent reflect the tendency for NCR data to signal only the most important recovery cue, while a rich span of think aloud data may allow several behaviour codings to be applied.

This point may be worth dwelling on: in the Ericsson & Simon model of verbal report, the think aloud informant ideally just thinks sequentially about her progress through the task and makes no attempt to select what to report. This

picture of informant task-processing may or may not reflect what actually happens, but think aloud and retrospective data gathered from solo-informants suggests that conscious 'editing' is limited. The NCR informant, on the other hand, is able to, and arguably has to, select among the impressions and inferences that accompanied her recovery those that seem reportable in a fairly concise way. (As there is no preset time-limit on NCR reporting phases, the most likely reason for their comparative brevity is that informants have said all they are able to tell or all they think worth telling.) Think aloud and NCR, then, then, arguably diverge not just along the axis of 'the time-lag between event and report that impacts on reliability' emphasised in Cohen 1997 and elsewhere, but also in terms of control of reported data. NCR may grant the informant more freedom, if you like, than is available to her think aloud counterpart, and in this respect may be seen as a half-way step to the AC procedure taken up in following chapter.

6.14 'Leading the witness' in post-task interviews?

I wish to discuss briefly here a fairly basic issue in comparing think aloud to NCR reporting. As noted in the case of Harumi, above, it is not always possible to schedule an interview immediately post-task, and if the informant is fatigued, or unhappy with her think aloud performance and/or her cloze success then even a genuinely post-task interview may be only marginally productive. I mentioned, above, the risk of biasing an informant's retrospections through 'leading' questions (which, it must be recalled, are in one party or the other's second language) and I must accept that I have fallen into this trap on a number of

occasions. When one reviews a think aloud informant's audio-protocol, there is usually time for only a single listening in her presence. It is only too easy, then, to form a hurried impression of what was 'meant' by a particular remark, silence, or intonation.

One naturally seeks to validate that impression with one or more open questions, but where an open question fails to elicit a useful response, 'Were you...?' or 'Did you...?' questions may be used to elicit the informant's interpretation of what was going on at that moment. The danger will be clear; if the informant cannot very clearly recall her processing, as it appears that many cannot, then a range of interpretations may seem equally plausible to her. Much the same may happen to the researcher: on transcribing think aloud data with my interview notes to hand it has sometimes struck me that my initial interpretation, sometimes 'confirmed' by the informant, no longer seems very plausible; indeed another and more likely interpretation altogether may have presented itself. The same danger must be presumed to exist in post-task interviews of NCR informants, although my experience suggests that it will be markedly less: to date I have not come upon an instance in which I have reinterpreted an NCR informant's protocol data in a way that conflicted with her post-task retrospections.

Doing without post-task data?

The thrust of these remarks is of course that the interaction of researcher intuitions with informant retrospections (or lack of same) is not always conducive to the production of valid and reliable data. But if we are to do without post-task

interviews then it is necessary to consider how far the 'on-task' data from a verbal report format allows one to code processing behaviours and events on that basis alone. This has proven problematic: in almost all cases, think aloud informants' responses in post-task interviews or to post-task questions were taken down in note form in the TL (i.e. translated from the informant's L1 as necessary) and condensed or paraphrased. Some codings could be applied solely or very largely on the basis of retrospective comments. These are identified in protocol extracts by underlining, and can readily be separated from events derived from task-concurrent data (cf. GL1 Claudia item (4): "[...] and then after the rules.. *against* foreign competitors had been waived international"; this, seen as uncodable through the think aloud data alone, was coded as 'Col' on the basis of the informant's post-task remark that 'national' and 'international' "belong together")

A second category of event consists of those which are so clearly identifiable from concurrent verbal report as to need no confirmation (cf. Claudia, item (3): "[...] because later in the sentence there's international" coded as 'SSf'.)

There is, however, a third category of event, in which codings were, with varying degrees of confidence, applied on the basis of task-concurrent data, and then confirmed by the informant's retrospective remarks. Given that, in the majority of cases, informants' post-task observations were taken down in note form, translated as necessary and condensed or paraphrased, it is not possible at this distance in time to say with confidence whether or not some of the events in this

intermediate class could have been safely derived from task-concurrent verbal data alone. This makes it unrealistic to aim, in the circumstances, at an accurate comparison of events codable only with, and without, recourse to post-task retrospection. Consistent recording of post-task interviews as well as the recovery task itself would likely have ameliorated the problem–although this would have required a second recording device each time. Had I been conscious at the time of the desirability of assessing the relative contributions of concurrent and post-task data, I might however (given informants' cooperation, and time enough) have been able to arrive at a much clearer assessment through note-taking alone.

To sum up., the importance of post-task information in providing insight into the processing of a cloze passage varies from one informant to another and (see below) with the reporting format. Some think aloud informants expressed themselves clearly enough that relatively few areas seemed to need clarification post-task, while others left rather more 'open' yet (cf. Harumi, above) could not offer much useful retrospective comment.

In the absence of an accurate overall comparison of relative contributions of on-task and post-task data we fall back on individual cases, and it may be useful to look at the role of post-task retrospection for those informants whose protocol data is shown in this chapter. The key points of my task-observation and post-task notes accompany the protocol data above. To take Claudia first, her post-task interview provided confirmation of five to six codings which I had tentatively applied while listening to her think aloud. More importantly, in five recoveries

(items (4); (6); (14); (22); (25) see protocol, above) or some 15% of items, the only codings which could be made were based solely on Claudia's post-task comments. (This is an illustration of the potential contribution of post-task retrospections even from a productive think aloud informant.) Post-task comments, in fact, cannot be relied on to compensate for limited task-concurrent reporting: asking an informant to retrospect about an item which she failed to verbalise about at the time is seldom productive, and for two reasons: (a) with very sparse protocol data there is little to stimulate informant recollection, and (b) there is little or nothing against which to validate any retrospections she may make. Poor concurrent reporters thus tend to be, in effect, weaker retrospectors.

When we look at the role of post-task comment by JL1 informant Yasuko, we see that she provides confirmation, post-task, for between three and five codings, with only one ('Col' in item (7 above)) made possible solely by her post-task comments. The major difference between Yasuko and Claudia in this respect appears to be that the former more reliably made mention of grammatical knowledge or cues. Harumi also offered some confirmation of two to three codings (items (3-4) and (8), above) and only the coding of items (22-24) seem to be based very largely on her retrospective remarks.

In the case of paired informants Anneke & Fred, and Mitsuo & Arisa, post-task confirmation tended to focus more on what we might see as more 'individualistic' behaviours such as translation (Anneke & Fred item (7), above), as well as on such aspects of informant interactions as whether informants were working

together or separately at a given point (Mitsuo & Arisa, items (1) – (6), above). The somewhat reduced role for post-task comment in paired think aloud reporting reflects the greater interpretability of the more structured and already 'externalised' concurrent verbal content, and (although, as discussed above, it is hard to cite accurate percentage figures) the coding of only a very few items seems to depend on retrospections.

Turning to NCR informants, Ryou is representative of this format's informants in that his concurrent report data requires little clarification, and no item codings depend solely on retrospective data. (Among NCR informants, no more than ca. of 1% of items were coded only on that basis.) Given the relatively higher interpretability of their reporting, post-task questions to NCR informants had more to do with confirmation of points than with clarification as such: Ryou, for example (see protocol, above) confirmed that in item (27) he had contemplated using the extratextual knowledge (that the ancient Olympic Games lasted six days) implied by his concurrent report, but that had opted to make a "safer" logical inference based on textual data. The clearer the initial verbal data, one might say, the greater the scope for 'safely' going beyond the surface features to get at subtler aspects of processing–of which Ryou's weighing of prior knowledge and textual information is a good example.

6.15 Conclusion

This chapter has presented think verbal report data in two variants, think aloud and NCR, and illustrated the kinds of data the former provided from GL1 and JL1 informants, and the latter from JL1. It was noted that the second format, NCR, appeared to produce markedly more interpretable data, and appeared to 'track' key events in cloze task-takers' processing of the passage to the extent that, for many purposes, the data gathered via an 'immediate retrospection' technique such as NCR may be adequate (cf. the detailed comparison of think aloud, NCR and AC products in Appendix 4.) Given the comparative clarity of NCR verbal data, the amount of time required for post-task interviewing was lower, as was that taken up by transcription and consultation about unclear points. While the low numbers of informants does not allow a firm conclusion to be drawn (but cf. the results of the 'hypothetical situation' opinion survey reported in chapter 8) I suggest on the basis of observation and post-task informant remarks that NCR informants found the cloze-plus-reporting task less stressful than their think aloud counterparts. In short, the advantages of NCR over think aloud in practical terms suggest that a researcher contemplating the use of a verbal report procedure might usefully make her initial trials with as the former and assess whether the data produced are adequate to her needs. If not, the potentially richer though unarguably more time-consuming alternative of think aloud is available.

It seems undeniable that, for all its advantages in terms of time and effort required by the researcher/analyst, and the apparent affective benefit to some informants,

NCR does carry the risk that data which might have been tracked by think aloud will be lost. I take up in chapter 8 the question of how salient a drawback this potential data loss may generally be, but acknowledge here that the decision as to which reporting format to employ will vary with the goals of data-collection, and the real-world constraints of time and facilities available. In the end, one can (or perhaps should) base one's evaluation of data-elicitation procedures on one's experience of actually using them. In the following chapter I turn to another attempt at getting inside cloze test-takers' heads by what was intended to be, from the researcher's point of view, a speedier and less labour-intensive alternative to verbal report, requiring informants to represent their own processing behaviours via a set of codes and graphic conventions, supplemented by written comment.

CHAPTER 7: ALTERNATIVES TO VERBAL REPORT

7.0 Introduction

So far, the data-elicitation procedures looked at have relied on the verbal reporting of processing events: in think aloud the informant is expected to keep up a more or less continuous stream of verbalization about her processing of each cloze item; in NCR she is asked to report at intervals about how she recovered each filler. The most salient events and processes in cloze recovery–though by no means all of the cognitive events identified in think aloud protocols–appear to survive in selective NCR reporting, and I have suggested that a choice of which procedure to employ will depend on the researcher's specific goals, the level of detail required, and the time available for transcription, etc.. The high demands made by the gathering and preparation of think aloud data are well known (Trollope 1995; Green 1998) and have already been outlined here. Although NCR places rather lower demands in time on the researcher (as well as on the informant) it still relies on access to recording and transcription equipment, and in the absence of suitable LL facilities this results in a more drawn-out data-gathering process, with (I claim) a greater attendant risk to security.

It is tempting, then, to look for a still faster means of getting at information about informants' processing of passages. In this chapter I look at the pros and cons of two means of accessing information about task processing behaviours. In the first of these, questionnaires, the collection of verbal data is optional, while the second ('annotated cloze' or AC) dispenses with verbal data entirely, in favour of the informant's self-coding of her own processing events from an *a priori* checklist. It

is useful, I think, to consider the two methodologies in tandem, for although in some senses (e.g. 'open-ness' or scope of response) they may differ, in others (such as the pre-selection of aspects of processing for attention) they may not be so far apart. I discuss first my application of questionnaire items in looking at aspects of cloze processing perhaps less amenable to tracking by other methods.

I then take up the development and use of the AC recording format, and what it can tell us about how cloze blanks get to be filled. One aspect of the AC task, that of graphic marking up on the informant's 'manuscript' appears to lend itself well to the recording of one processing behaviour, translation, whose role was (cf. chapter 6) not always readily traceable via think aloud. I thus look in more detail at AC's potential as a means of getting at the role played by L1 translation in cloze tasks. I go on to compare AC with think aloud and NCR verbal reporting formats, to consider the possibility that reporting condition may affect cloze score, and to look at the effectiveness of the three conditions in terms of how comprehensively they elicit data about the processing of items, and in what depth.

7.1 Questionnaires in data-elicitation

A number of researchers (cf. Purpura, 1997; Cohen 1998) have proposed the use of questionnaires—typically administered as soon as possible following the task—as a means of getting at what test-takers are doing as they process the test material. Leaving aside for now the question of whether some of these proposed instruments are better characterised as 'checklists', I illustrate now problems with the use of questionnaires with reference to my own brief trial of an on-task

questionnaire. This exploration came about simply because I wanted to see what kind of information respondents might volunteer in response to both 'open' and less open questions. Given the 'immediate' and unfiltered nature of much think aloud data, I did not expect that responses to questionnaire items would reveal anything which had not already been found in that reporting format. These responses might, however be more readily interpretable, and might well capture more economically capture some of the same kinds of data that verbal report seemed to provide. Might questionnaires provide triangulatory data to confirm or disconfirm information gleaned from verbal report? Might I (as was the case in the realization that think aloud pair informants may consciously withhold relevant information or insight which they feel their partner already possesses) uncover something I had previously overlooked?

'Open' questions, or at least responses to them, are something of a double-edged sword. On the one hand (an advantage shared by verbal report) they do not constrain informants to answer in predefined ways, and hence may record a wider range of behaviours and stimulate new insights. On the other hand, informants' responses may (again, much as in verbal report) diverge greatly in terms of how much information they contain and how readily they can be interpreted. The more 'open' the question, moreover, the greater the role of productive expression or articulation is likely to be. Open questions may also require extensive trialling and revision (Cohen & Manion 1994) in order to ensure that they are consistently understood by informants in the ways intended. I thus opted to use an open

question only where it seemed desirable to constrain the respondent's replies as little as possible. For those aspects of cloze processing about which a more restricted set of information was desired, selected-response or 'scalar' items were taken to be adequate.

The obvious target for an open question was to ask how respondents had arrived at their chosen filler, while restricted response items could look at aspects of processing that might have been traceable in think aloud only via interruptive questioning (something which the Ericsson & Simon 1984/1993 model still appears to frown on, and which NCR reporting did not reliably track. Informants' 'language(s) of processing' (which, cf. Cohen 1998 need not be the same as the language(s) of reporting) seemed like one worthwhile target for a restricted response question. Although respondents' perceptions of item difficulty were tracked with moderate consistency in verbal report data, I wondered if questionnaire items would more reliably and/or more economically provide the same information. Nevo 1989 reports what appears to be a successful combination of 'checklist' procedures and questionnaire items, and in the light of her remarks it seemed worthwhile to explore the use of the latter further.

7.2 Format of the questionnaire task and session

In a language-lab LL setting ten 'fresh' Japanese L1 informants were given a written questionnaire on which questions *X*, *Y*, and *Z* (These letters are used to stand in here for the actual deletion numbers; these were omitted from the printed master version and added *ad hoc* by hand. The actual questions appear below.)

were repeated as necessary. To minimise any comprehension difficulties stemming from the rather brief task orientation that preceded the task, all rubrics were provided in both TL and L1, and L1 translations of the questions were shown to informants beforehand. The orientation included a short 6-item practice passage (adapted from FUNCHAL) while the target passage itself was extracted from the longer text REFUGEES. This is shown below with deleted content underlined. The task was kept fairly short (11 deletions) simply to avoid difficulties of scheduling.

REFUGEES

The world's refugee population is on the rise. Some 12 million people have been granted "refugee" status. This gives them certain protections and rights. But nations are slowly closing their (1) doors to asylum seekers. Resettlement officials say (2) the world community appears unable-or unwilling-(3) to mitigate the underlying causes of refugee (4) movements. Thus, they see a need for (5) greater public understanding of refugee issues.

Twenty (6) years ago there were 1.5 million refugees. (7) Today there are over 25 million, according (8) to the office of the United Nations (9) High Commissioner for Refugees (UNHCR). The steady rise (10) in the refugee population is mainly the (11) result of international conflicts.

For timetabling reasons the actual session had to be kept fairly brief, so that a short and not overly-challenging passage seemed appropriate. (The number of items the passage contains is too small for a valid cloze passage (cf. chapter 2; Alderson 2000) but this was not intended to be a test as such.) The rationale for presenting a written questionnaire in a LL setting was that this would allow respondents to answer in their preferred mode—which might or might not remain the same throughout. Although answering a questionnaire may not closely

resemble a think aloud task, I was concerned that respondents (who had in fact been invited to this session mainly because they would soon graduate from the courses on which I was teaching, and hence be unavailable for later use as informants) included some 'quiet types' who, though able, often appeared reluctant to speak in class activities.

All respondents thus had the option of writing, and/or of recording their responses at any time at the press of a button. Informants were instructed either to speak or to write their answers (in the blanks below each question) or to mix both modes as they liked, at the appropriate points in their processing. It was anticipated that respondents would be most likely to respond orally to Question X, as this was likely to elicit the most detailed information–which written responses were less likely to adequately record. I hoped, however, that respondents might also add oral comments to their written or graphic responses to the other items. As was the convention in my classroom lessons (and this data-elicitation session was intended to present a didactic veneer, as it were) informants were told that English should be used for things they could readily express in that language, but that Japanese should be used to respond whenever this seemed necessary. Written responses were to be made in ink (with instructions to score through changes and write in revisions or corrections above) in order to preserve a record of any changes made. The recording equipment was also set up in such a way that only I could rewind tapes. If an informant wished to revise a comment, she simply had to record it again in its revised form. The questions asked are shown below, and are labelled

as *X*, *Y* and *Z* rather than with numbers because each could of course appear at any point in the questionnaire. Space for written comments was left below each question. This made the questionnaire seem rather lengthy, but also low in 'density'; informants' comments following the task suggest that this format was viewed positively.

Question X

"How exactly did you decide which word to put in blank number _____? Give as much information as you can."

Question Ya

"As you worked out the best choice of word in blank number _____, were you thinking

(a) basically in Japanese

(b) basically in English, or

(c) in a mixture of both languages?

If (c), which language were you using more–(E)nglish or (J)apanese?

Question Yb

On the line below, mark an 'X' to show the language or languages you were thinking in as you worked out the best choice of word for blank number ___ .

|-------------------------|-------------------------|

E only 50:50EJ J only

Question Za

On the line below, mark an 'X' to show how difficult you think it was to work out the best choice of word for blank number _____ .

|-------|-----------------|-----------------|-------|

fairly easy fairly difficult

Question Zb

Was it (1) very easy (2) fairly easy (3) fairly difficult, or (4) very difficult to find a word to fit blank number _____?

Figure 7.1: format of questionnaire to accompany a cloze passage

The difference in phrasing between questions forms Za and Zb (one 'end' of Za is

labelled 'fairly easy' while the first choice in Zb is 'very easy') is not accidental. I had noted in classroom applications of 'mark an X' response lines that Japanese L1 respondents appeared to avoid marking extremes labelled 'very easy'. By labelling the response line as I did I hoped to see whether respondents would be more willing to, in effect, choose 'very easy' by marking the line to the left of the label 'fairly easy' (glossed as *wari to kantan* in the Japanese version of the rubric) in the absence of the 'very easy' label. This information, I felt, might be of value in classroom activities and further data-elicitation. Any or all of the questions above might be directed at the same deletion, or at different deletions. I did not have enough informants in this session to ask each question for every deletion , but I was able to compare two forms of each question in enough instances to get some idea of whether one or the other form was more productive:

Respondent 1	Respondent 2	Respondent 3	Respondent 4
Question X	Question X	Question X	Question X
Question Ya	Question Yb	Question Ya	Question Yb
Question Za	Question Zb	Question Zb	Question Za

Figure 7.2 (partial) allocation of question forms to respondents

The rationale for phrasing question X in that form was that it would be minimally restrictive of content, while still focusing the informant's attention on how she chose the filler word. The rationale for having alternate forms of Y and Z was that I wanted to compare responses across the 'multiple-choice' selective format, so closely associated with tests in Japanese educational culture, and the Likert scale-like graphic format which informants had used quite extensively in their

English classes (e.g. in surveys of class opinion) and with which they were thus familiar. Answering questions such as the above may require a good deal of time and thought on the informant's part, and may (see section 7.5, below) be more disruptive of processing than other verbal report procedures. Even where informants can answer the question reasonably fully, however, their answers may be phrased in disparate ways, so that it is hard to be sure whether (oral) responses (to question X) such as those shown in [A] and [B]m below, can equally be taken as evidence that recall of 'formal learning' was a central factor in recovery.

[A] "I think I could remember [the word]... maybe I learned it. [LAUGHS] WELL...MAYBE" (JL1 Miyuki)

[B] "MAYBE I LEARNED [THE WORD] IN SCHOOL... DUNNO [*SAA*], I KNOW IT FROM SOMEWHERE OR OTHER" (JL1 Hideki)

While informant [A] could perhaps have been pushed to provide more detail, the tone of her response suggested that she would not have been able to say much more with confidence. [B], likewise, might have narrowed down the context (secondary school?; language school?) in which he acquired the word, but the L1 choices of *"saa"* ('dunno') and *"dokodoko"* ('somewhere or other') do not imply that he had much more to tell.

I have to add two caveats here. I subsequently became aware that in the manner in which I trialled my written questionnaire as a data-elicitation device I may unintentionally have undermined its importance in respondents' eyes. One way in which this may have occurred is that, through an administrative error on my part, only a subset of theoretically available informants were actually asked to

participate. For timing reasons, moreover, both the practice and target cloze passages and the session itself (ca 35 minutes including the orientation) were brief compared to previous think aloud data-elicitation sessions. Moreover, due to scheduling difficulties I had not been able to schedule post-task interviews. I must assume that at least some of the participants in the session outlined here were aware that previous sessions had been both longer and 'multi-stage', and that the contrast may have created an impression that this data-elicitation session was somehow less important.

The second caveat may be more important, and has to do with the possibility that exposure to questionnaire items will influence respondents' subsequent processing behaviour. This influence may operate at a more-or-less conscious level, such as when an individual who has been taught to maximise her use of the TL tries to use it in response to a question about language of processing, when in other circumstances she might have used her L1) but 'unconscious' influence cannot be ruled out. With hindsight, it would have been better to have used some variant of the 'envelope-and- single-question-item' presentation outlined below.

Responses to the practice questionnaire passage: overall brevity

Whatever the reasons, written responses to question X in the practice passage were seldom very enlightening. The same was true of those oral responses I had time to audit and note down. In both modes, many answers were sparse:

"I translated." (Oral)

"[I] KNEW THIS EXPRESSION." (Written)

"I JUST KNEW THE ANSWER." (Oral)

"I knew some words and chose the best ONE." (Written)

and rather more than expected were in the informants' L1, including things which they very definitely knew how to express in the TL in a classroom context.

Language of processing & item difficulty

For question Ya almost all informants chose option (c) plus (J) indicating that they had used both languages in processing, with Japanese predominating. The 'Xs' of those who used question form Yb were concentrated approximately 60% along the line from the left:

```
|------------------------|------XXXX----------|
E only                 50:50EJ              J only
```

Figure 7.3: approximate distribution of responses to question 2b

In a sense, this echoes a finding of an earlier study of 'language of processing' in cloze, conducted with German L1 undergraduates. Respondents in that study were given, at unpredictable intervals, a forced choice in which they had to indicate whether they were (or had just been) thinking in their L1 or in the TL; n some 40% of cases no selection was made. I would speculate that respondents may find it hard to distinguish between the language used in 'silent reading' or 'reading aloud' of text content (presumably the TL, except where translation or L1 paraphrasing is taking place) and the language used in thinking about that content. To some extent, option (c) here may also represent an inability or reluctance to make the choice of language of processing. Difficulty levels were often left

unmarked in the practice passage, but the fact that auxiliaries deleted from verb phrases were sometimes marked as (3), i.e. 'quite difficult', is rather at odds with the often 'immediate' recovery, typically by recourse to knowledge of grammatical rules (where this is in fact recorded) in think aloud protocols. Think aloud informants' retrospective comments had also indicated that such deletions did not, on the whole, present much challenge.

7.3 Towards better reporting & responses in the elicitation session

When almost all informants had completed the practice passage, I collected their response sheets, which–once shuffled–were effectively anonymous. While participants took a short break I isolated some responses for comment. I pointed out to the reassembled group that a response like "Knew this (word/phrase)" might be more valuable if the informant had tried to recall where she had learned it: in school, perhaps, or in a recent lesson. (This information, I was aware, might not be available in very many cases, but my aim was to stimulate informants to think in greater depth, and to say more, about the knowledge they were applying.) Similarly, it would be helpful to try to give some idea of how or why one word was chosen over others. Was there some clue in the passage that made this word preferable? Did it connect to something the informant already knew? Did it just 'sound better'? I also pointed out the inconsistency (*itchi shinai*) between—as at least one respondent had done—filling in a deletion correctly, specifically stating that she recalled the context phrase from a recent classroom reading activity, and then marking the item as 'very difficult' to recover. I emphasised, too, that I was

interested in respondents' authentic perceptions of how they arrived at an answer, and of how difficult it had been to do so. The questionnaires, I emphasised, would be anonymous, and the task could have no effect on any grade. One informant's stated desire to later be told her 'score' was echoed by others – evidence again, perhaps, for my suggested overlap between tests and 'test like' data-elicitation tasks. I thus told the group to each choose a 'pen name' and write it on their papers. I would score the tasks, add my comments, and later make the papers available in individual envelopes.

We then moved on processing of the actual target cloze passage (REFUGEES, adapted). The responses gathered were—perhaps as a result of my earlier comments on the practice task results—on the whole longer and more detailed, and a little more productive of insights into processing. No marked difference was noted in this regard between written and spoken responses. In general the meaning of comments was clear: "I translated this part to get the meaning."; "I chose the word that sounds best." A few 'stripped down' written responses in the L1 had to be expanded, e.g. "*koukou de* (in high school)" was taken to mean "I learned this [phrase] in high school."

Many responses to Question X were brief or even monosyllabic. Questions targeting 'grammatical' deletions (2), (3) might consist simply of comments like "Isn't it [a question of] grammar? *(bunpo ka na)*";"This is English grammar word" [*sic*]; "grammar". Question X responses to (if the 'UNHCR' clue is counted) 'extratextual' deletions such as (9) [HIGH] could be more rewarding: "Everybody

knows that", which implies that the respondent had little trouble filling the blank and thought it easy for anyone–which, at a SEMAC score of 60%, it was not. The deletions within the target passage are shown below, along with (in parentheses) alternative fillers deemed SEMAC correct by at least two of four TL NS consultants and myself. The respondents' answers to Question X are paraphrased for economy, and in terms of the behaviours already identified in verbal report data. (In this question. each respondent was asked to comment on either two or three items.):

(1) open their doors (door; gates)
Knew the Phrase/Collocation/Idiom 'open/close the door to': 2 of 3; Extratextual knowledge/Logic (only doors close to keep people out): 1 (all SEMAC)
(2) the world community
Used grammatical knowledge: 3 of 3 (all SEMAC)
(3) unwilling to mitigate
Used grammatical knowledge: 3 of 3 (all SEMAC)
(4) refugee movements (flight)
Apparent use of collocation: 'refugee problems'; 'refugee people'; 1 blank (2 of 3 non SEMAC fillers)
(5) need for greater understanding (more; better; wider; deeper)
Previous sentence implies world is unwilling to help: 2 of 3; Extratextual knowledge/Logic (Resettlement officials would not ask for *less* understanding): 1 (all SEMAC)
(6) Twenty years ago
Knew the phrase/collocation 'X years ago': 2 of 2 (both SEMAC)
(7) Today (Now; Currently)
Contrast with time reference in previous sentence: 3 of 3 (all SEMAC)
(8) according to
Knew the Phrase/Collocation/Idiom 'according to': 2 of 2 (both SEMAC)

(9) High Commissioner (Chief)

Extratextual knowledge: 2 of 3 (both SEMAC) 1 blank

(10) rise in the refugee

Knew the Phrase/Collocation/Idiom 'rise in': 1 of 2; used grammatical knowledge: 1 of 2 (not SEMAC)

(11) the result of

Knew the Phrase/Collocation/Idiom 'the result of': 2 of 3 (both SEMAC); used grammatical knowledge: 1 of 3 (not SEMAC)

Figure 7.4: Paraphrased responses to Question X, above.

While the sample size is clearly too small to draw more than inferences regarding behaviour in Question X, it is interesting that in item (5) one informant seems to have used extratextual knowledge which was not strictly required. This event appeared on audiotape and appeared to have been a more or less immediate, initial response.) Also of note is that the same respondent claimed to have used grammatical knowledge in items (10) and (11), and got the filler wrong in both cases ('of' in (10); ('cause' in (11).) The two who claimed to have known the phrase or collocation appear to have actually done so, as their fillers were correct. (Little can be read into this, however, as informants do not consistently distinguish 'phrasal' and 'grammatical' knowledge in relation to noun/verb-preposition links; a wider study might establish whether phrasal/collocational knowledge represents a more 'concrete' basis for recovery.) Overall, the responses to question X were in line with the processing events recorded by think aloud and NCR.

Responses to Question Ya (language of processing) differed from those in the practice session. Eight out of ten informants chose (c) while two chose (b). Six of

those who chose (c) also claimed to have mainly been using English, and two their L1. Responses on the line in Question Yb were grouped around the (from the left) 35-50% range in favour of English. Overall this suggests that respondents were using the TL slightly more in this task than in the earlier task. This may have been because the second passage was (in their view) slightly easier, or because they were more familiar with the task. I do not think that my critique of practice session responses influenced subsequent choice of language, although this cannot be ruled out. Again, questionnaire reponses do not lend themselves to fine distinctions such as that between the use of the TL in reading passage content and in thinking about how to fill the blank. Individual informants appeared to answer this question fairly consistently across the deletions to which it was applied, although two informants' line marks for Yb in relation to item 4 were at around 60% in favour of Japanese. This is the one item which attracted no SEMAC fillers, and was rated most difficult by respondents. Responses to Questions Za and Zb (item difficulty) were consistent. The mean scale ratings for items (using responses to question Zb) were:

(1) 2.3 (2) 1.5 (3) 2.3 (4) 4.3 (5) 1.7 (6) 1 (7) 1.5 (8) 1.3 (9) 3.3 (10) 2.5 (11) 2.7

Although some items were coded as (1) ('very easy') in question Zb, no 'x' marks appeared at the 'extremely easy' end of the line in question Za, suggesting that the range of 'linear' responses to a question about item difficulty may be artificially constrained in this reporting format. The information about item difficulty gathered via question Zb is clearer than that available through think

aloud/retrospection, and was more consistently recorded at 90% than was true of NCR informants (ca. 66%.) Against this, the questionnaire passage was considerably shorter and the reporting format arguable less distracting than the obligation in NCR to verbalize at each recovery; the rate of response to a question about difficulty might be expected to decline with a more challenging passage. Returning to question Y, the information gained about language of processing is also clearer than that offered by think aloud and retrospective data, supporting Nevo's 1989 integration of questions into a checklist task. I discuss below the timing and presentation of questions about aspects of processing.

7.4 Delayed responses to questions

One aspect of questionnaire respondents' behaviour (which I had not seen reported elsewhere) attracted my attention as something which may require explicit and specific instruction *if* it is seen as undesirable. I noted that a number of informants (inevitably, I could not observe all of them, all of the time) appeared to delay responding to the questions about a recovery until well after they had put out at least some processing effort on that deletion. Observed 'focused attention signals' (such as tapping the point of the pencil repeatedly on a blank, reading aloud a span of text leading up to or surrounding the deletion, 'sounding out' candidate fillers, or of course actually filling in the blank) made it possible to follow approximately, and for at least some of the time, where in the passage those informants in clearest view were focused at a given point.

Given that respondents speaking their answers into their headsets often had their faces partly obscured by hair or arms, this delay effect was clearest in observed written responses. I concentrated my observations on these respondents (or at least those closest to me) in the expectation that I would be able to decipher much of what 'oral' respondents had been doing from their audio-recordings. (While three respondents did roughly balance written and oral responses, the remainder displayed a fairly clear preference for one more or the other as their main mode of communicating information.) This expectation, it may be worth mentioning here, turned out to be a mistake. In recording previous think aloud sessions I had used whenever possible tabletop or clip-on 'lapel' microphones; these picked up the sounds of writing far more effectively than was the case with the small LL headset microphones in use on this occasion. I could, by and large, thus be sure of when an oral respondent in this task was writing in a filler answer only if she made some utterance to that effect. Fortunately, informants quite often flagged the insertion of a filler or answering of a question with a repetition of the question, or an expression such as "Right (*jaa*)"; "I'll write [it] (*kaite okou*)" or "That's done (*dekita*)."

On these assessments of informants' foci of attention, it was by no means unusual for the answering of a questionnaire item to lag one or two blanks behind the processing of its related deletion. Briefly questioned post-task, one informant who had shown consistent and marked time-lag between processing and writing seemed unaware of having done so. Others were conscious of having delayed

responding to questions. One told me that she was aware of having delayed writing in passage fillers in order to see if a better idea came up. The implications of this time-delay are interesting, for in cases where the informant has already filled in a word, what has she to gain by delaying her report of how she arrived at that word? It is intuitively plausible (and this is the sense I had during observation of informants working on the task) that some respondents were focused on the passage-task to the detriment of their responses to the questionnaire itself. This might be overcome by more thorough orientation and practice. Alternatively, the design of the task may not have been conducive to the idea of reporting in writing and later correcting—even though I explicitly described how to make revisions or corrections (score once through the part to be amended, and write in the revision above) and why (preservation of steps in processing) they should be made in that way.

Perhaps the most plausible explanation, however, is that (in the absence of any checklist of options) informants simply wanted more time to optimise their responses. I did not attempt a realistic self-as-subject study of the questionnaire format, but a near bilingual Japanese consultant to whom I subsequently set the task (though using a 29-deletion version of the OLYMPICS passage) also exhibited a time-lag between working on, or even filling in a blank and (in this case orally) responding to the question of how she had arrived at the filler. This informant claimed afterwards to have been aware of delaying her responses in order to be more sure in her own mind of her processing, and to be better able to

express this clearly in her (mixed L1 and TL) responses. She had, however, no answer to the question of why she had not written down responses such as 'level of difficulty' more quickly, as these would appear to be open to swifter decision. That said, informant's perceptions of a cloze blank's difficulty must in large part depend on whether or not they felt able to fill it acceptably, and delay even in rating an item's level of challenge might be interpreted as reflecting an ongoing effort to assess a filler in relation to later passage content and/or subsequent fillers. For other aspects of the response task, it would not be surprising if, charged with constructing an informative answer to a question she would perhaps normally never need to consider, an informant might delay her response until she felt satisfied with it.

Reducing the risk of bias when questionnaires are used

Whatever the factors behind delayed responses to questionnaire items may be, the practice may be problematic in terms of the classical model of verbal report, which still places considerable value on immediacy of reporting. In Cohen's terms (Cohen 1984; 1998) questionnaire responses might be categorised as something between 'self-observations' and 'self-reports', with a consequently greater risk (in that author's model of processing) that responses will be reconstructions of processing rather than reflections of what actually occurred. The delay noted between processing events and their reporting might to some extent be 'designed out' by a revised questionnaire format and better respondent training. As we shall see in the following sections, the perceived difficulty of individual cloze items

may be fairly readily tracked by the incorporation of difficulty rating scales in the AC task. An aspect of cloze processing in which the AC format seems weaker, however, is in the recording of processing events which are not tied strictly to *individual* blanks. Questionnaire items might usefully be applied to such areas as the application of any pre-recovery activities such as pre-reading; the balance of languages of processing in use at a given point or points in the task; the use or not of revision, etc. Some care is called for in the use of question items like these, for while it is clearly preferable to 'set' questions as close as possible in time to the events they focus on, setting them prior to those events is problematic: it is known (Cohen & Manion 1994) that questionnaire items can bias subsequent behaviour.

A useful technique which I have recently put to didactic/classroom research ends may be worth outlining here. Question items are typed on slips of paper and either numbered on the back or placed within small numbered envelopes which are distributed along with the task, and participants are instructed to turn over a slip/open an envelope and answer the question only when cued to do so by the teacher/researcher, or at a given point in their completion of the task. In an investigation of cloze processing, informants could be asked to open the envelope containing a question, such as that below, about pre-recovery activities, as soon as they had filled, say, blank (3):

"How much of the passage did you read before you started to fill the blanks?"

(a) very little (b) the first paragraph (c) all of it

If you chose (b) or (c), were you (please circle the best answer) just reading / reading and filling some blanks at the same time?

Figure 7.5: An example single-question-item

The delay in seeing the question obviates any risk of bias to the early stages of recovery, yet still gathers responses while the information is still available. I have found this technique to be very successful indeed in terms of completion rates (95% and 97% over two applications) and ascribe this to the twin factors of curiosity and the difficulty of overlooking the slips or envelopes. Through this procedure I can report that just over 64% of students (n=25) on a didactic lexical inferencing task reported pre-reading at least the first half of the passage before beginning to infer meanings (The task carried no instruction to pre-read.) but that of the same group only 28% reported doing so *before* beginning to fill blanks on a didactic cloze task two weeks later. (The possible anticipation of a question relating to pre-reading of the cloze passage does not appear to have artificially stimulated this behaviour on the task to any great extent, suggesting that the two week interval was long enough.)

I conclude this discussion of the application of questionnaires with this proposal for better presentation of question items, while acknowledging that my exploration of open and selected-response questionnaire items did not shed much new light on informants' cloze processing. The overall level of detail in the data gathered perhaps most closely resembles that of NCR, and questionnaires are only

likely to be markedly more economical of researcher time than 'structured' verbal report (cf. chapter 6) if responses are written, and of course interpretable, rather than oral. All in all, questionnaires may repay the necessary investment of time in construction and iterative refinement in larger-scale research contexts–i.e. in terms of rather higher informant numbers than those available to me in the studies discussed here. The following section of this chapter takes up the development and application of a 'checklist' procedure also aimed at more efficient gathering of data relating to how cloze blanks come to be filled.

7.5 A selected-response 'questionnaire'?

Asking informants to write down (as in questionnaire item 'X') or relate (as in verbal reporting) how they arrive at a cloze filler is, in a sense, asking them to construct responses about constructing responses, and one could plausibly argue that a selected response format would impose less additional burden on processing. One such format, multiple-choice, is moreover by far the most familiar testing mode for Japanese students at least. I contemplated investigating the use of a multiple-choice format questionnaire, but in the end did not do so. Given that AC, a task format whose foundation data had already been gathered, was itself a selected-response procedure—and, I hoped, as heuristically productive as a multiple-choice questionnaire was likely to be—the anticipated return of information did not seem to justify the additional effort.

The desirability of a 'better' means of gathering data about cloze processing became clear as I realised how much work would be involved in acquiring data

from a respectably-sized sample of informants, and how hard it might be to entice suitable informants in adequate numbers (for as earlier solo-condition think aloud informants seemed in some instances to have been reluctant to recommend participation to their friends, the supply of volunteers began to shrink. (The label 'snowball sampling' may be accurate.) A small-scale trial of a selected-response format had been carried out with GL1 informants, using a restricted set of codings, and this had proved productive enough to persuade me that the instrument merited further development. The real impetus for development came, however, when I returned to Japan and found, in the course of my attempts to gather verbal report data from JL1informants, that the proportion of 'low-verbalisers' in that informant pool appeared to far exceed that encountered in Germany.

While the NCR format reporting procedure (see chapter 6) had proven useful in reducing the researcher workload in gathering verbal report data, as well as appearing to ease the task of reporting for informants–though as the NCR informant pool may have been slightly more at ease with English than their think aloud counterparts this is no more than tentative suggestion–it did not appear to improve these low-verbalisers' task performance to any great extent. I take issue with those with those who suggest that individuals who cannot adequately report on their own processing of a task should simply be removed from the informant pool. In my own experience, these individuals were predominantly male, and given the gender composition of my language classes the pool of male potential informants was fairly small to begin with. An alternative elicitation instrument

which did not rely on verbal report might, I anticipated, be of value in allowing informants unsuited to verbal reporting to effectively contribute to research into task-processing behaviour on an individual basis. (As briefly discussed in chapters 5 and 6, pair-condition think aloud may be more productive with informants unsuited to solo thank aloud, but not all were keen to work with a partner. Pairing informants at different levels of proficiency can lead to discomfort or even discord (cf. Haastrup 1991) so that the pool of potential partners may be further reduced.)

Although 'checklist' response selections have been used in the investigation of task processing (Nevo 1989; Allan 1992b, 1995) and are thought by some researchers (A.D. Cohen, pers.comm.) to be a valid part of the researcher's armoury of methods, a number of issues present themselves. One fairly crucial question must be: where do the checklist items come from? Nevo 1989 based her checklist on "test-taking strategies described in the literature and on *personal intuitions* as to possible strategies which respondents might select." (Nevo op.cit.:204, my italics) respondents were likely to select." The first of these criteria would certainly be justified if the research contexts in the literature were a reasonable match (in terms of task, pedagogical background, etc.) for the multiple-choice reading test used by Nevo herself, but the mention of personal intuition as a criterion for checklist items may raise questions, especially as some of the fifteen 'strategies' on Nevo's checklist "were assumed a priori to be contributory and some non contributory" (ibid.) apparently also according to the researcher's personal intuition.. Allan 1995 has raised interesting questions about

the use of checklists, such as whether respondents clearly understand the categories available, and can apply these with reasonable consistency and accuracy to their own processing behaviours. Allan's response appears to be a qualified 'yes', with the caveat (as I interpret his remarks) that the use of checklists is not yet an established methodology, and that potential pitfalls may remain to be discovered.

I opted to obtain the items for my own checklist of processing operations simply by taking over the set of events isolated in the think aloud data of GL1 and JL1 informants. It would be naïve to think that intuition had played no part in the creation of this set, but my intuitions had, I hoped, been to some extent validated through their application to informant data. Moreover, as I discuss below, the set of codings applied in think aloud was intended to be used by AC informants as a basis on which to construct their own edited or personalised repertoires or palettes of processing operations, rather than as a fixed set to be applied in a similar way by all.

7.6 AC Terminology

AC is burdened by a number of descriptive terms, only some of which have already appeared: *Codes* or *codings* refers to the behaviour or event designators which informants (rather than the researcher, as in think aloud) are asked to set down on their task sheets, or *manuscripts* (see below). The *codeset* is the full list of task *a priori* behaviour codings isolated from think aloud protocols and made available to informants (see Appendix 2) and from which informants are invited to

select those codings which best seem to tally with their own processing of the task. These individual working repertoires of codings I have labeled *palettes*. . The term *manuscript* refers to the paper task sheet physically completed and marked up by informants during the task. *Annotations* refers to written remarks, or graphic representations made by informants on the manuscript, whether in the spaces provided or elsewhere. These include such devices as the circling or boxing of spans of passage text, the drawing of lines or arrows linking passage elements, the underlining of translated passage content, etc. as well as the addition of written elaborative comments on the manuscript itself.

PTQ refers to the post-task questioning of some AC informants immediately following completion of the task. This is distinguished from the more formal and detailed post-task interview carried out with think aloud informants. The idea behind holding brief and informal post-task discussions, *ad hoc*, with AC informants stems from the intention that AC should be a comparatively speedy procedure, and place a significantly lower demand on researcher time than think aloud. This goal is seen as overriding that of maximizing the information elicited, so that while it might be possible to glean more information via extended post-task interviews with AC informants, this would rather defeat the purpose of seeking an efficient alternative to think aloud. There is some evidence from PTQs (see below) in fact, that informants may not have much additional information to offer. Additional insight has also been gained pre-task, as it were, from the necessarily lengthier orientation sessions held prior to AC tasks, and from

participants' comments in the small-group 'validation session' (see below) that was scheduled.

7.7 The AC informant's tasks

In an AC data-elicitation session, the informant is asked to process the cloze passage as far as possible in the way(s) she would approach an actual test. Instead of thinking aloud as she works on the task, the informant has to categorise her own conscious processing operations in terms of a set of codings from which she selects the most appropriate items. In addition, she is asked to add to her manuscript any written or graphic annotations that might further elucidate her processing behaviour. As can be seen from the sample manuscript extract shown in Figure 7.6. below, a separate writing space is provided for every deletion; this is 'open ended' in that the margin extending to the left or right may also be used for annotations. In early versions of the AC task, codings were marked on a separate sheet, but this was later abandoned in favour of a single manuscript as the sole recording instrument, with the codeset provided on a separate sheet, and later still also on a set of individual cards as described below.

The rationale for this format was that the chosen layout of the manuscript page 'boxes' provided a visible reminder that information should be provided for every deletion, replacing the conventional reminder in think alouds to "Keep talking." My hope was that, by minimizing the need for researcher input during the session, it would be possible to (1) gather data from more than a group of informants at the same time, and (2) to free up time for systematic observation of informants at

work. The number scales shown in the figure represent a scale which may optionally be incorporated into the manuscript. This scale has been used to allow informants to rate individual items for difficulty, or to rate their confidence in their chosen filler. (Formats with two scales per item were trialled, thus allowing both of these dimensions to be recorded, but informants found them confusing and completion rates declined markedly.)

7.8 Metacognition & analytic privilege

As outlined above, informants are expected to provide at least two types of information during the AC session: 'codings' of their conscious processing operations, and written comments or graphic elaborations about their processing. The first of these, coding, may be the most controversial in terms of the Ericsson & Simon 1984/1993 model of verbal report. As already noted in chapter 6, Ericsson & Simon's model emphasises immediacy of verbalization: the more closely an informant's output follows her on-line, moment-by-moment processing of the task, the more reliable it is held to be. Metacognition or 'meta-comment' (Ericsson & Simon 1993) on the informant's part about her task processing is seen as unreliable, and the categorizing of processing operations is the job of the researcher or (op.cit.:270) some "automated system", and not of the informant herself. The task of recording one's own mental operations in the form of preset categories is undoubtedly a metacognitive one, and hence highly suspect in the Ericsson & Simon model. As I think will be clear from the think aloud and NCR data cited in chapters 5 and 6, however, metacognitions appear to turn up

spontaneously and unpredictably in the midst of other aspects of verbal reports of cloze processing (cf. Haaastrup 1991 on lexical inferencing.) Examples are not hard to find: here JL1 NCR informant Ryou *evaluates a filler*, speculates about a TL construction, glosses this in his L1 , and finally *comments on the nature of the task*:

"I put *of* here but it looks funny.. of varied ability.. don't you say of *various* abilities?.. VARIOUS ABILITIES [*NOURYOKU*] RIGHT? WELL I ONLY HAVE TO PUT A WORD IN .. so.. of varied ability.."

while below JL1 think aloud informant Yasuko refers to a 'regularity' in her cognitions about an aspect of the TL, namely its article system (for Ericsson & Simon 1993:311 this would be a "clear instance of a meta-comment"):
".. oh I don't know about *a* and *the*.."

My suggestion is that too many protocols contain informant meta-comments about aspects of the task or their own performance for these to be discounted in verbal reports of language processing: it seems almost beyond argument that the observation made by Yasuko, above, might be of value in explaining or even predicting her (un)successful recovery of other blanks from which an article had been deleted. Asking informants to code their own cognitive operations and processing behaviours may appear to go against the notion of the researcher's analytic privilege or responsibility, but it is not without precedent. Cohen 1980a, for example, mentions a proposed investigative procedure in which comprehension exercises would be followed up with a type of questionnaire to elicit (*a priori*?) the operations used in the recovery of unknown lexis, and similar suggestions have been discussed (Purpura; pers.comm.) by other researchers.

Although I do not follow Ericsson & Simon in largely discounting post-task responses as a potentially useful source of data, I accept that essentially task-concurrent informant data optimal; on this criterion data gathered via the AC task may be preferable to that gleaned through follow-up questionnaire-type tasks. The AC coding task is anyway far from entirely in the hands of informants, for they are working from a set of coding options derived from previously-gathered verbal report data.

7.9 Aspects of verbal report taken into AC: passage, categories, and caution about modelling

As AC is simply another instrument designed to elicit information to compare with that gained via think aloud, and with the same view to examining informants' cloze processing behaviours, it was clearly desirable to use either the same or a truly analogous task-passage. As explained above, I could not adequately explain why the OLYMPICS passage appeared to stimulate more verbal reporting than other, apparently very similar, passages, so that even though verbal output was not a criterion in AC it seemed worthwhile to recycle the same stimulus passage. The passage was presented in almost the same format, with equal-length, individually numbered fixed-ratio blanks, and the same (ideally, for clarity) one-deletion-per-line layout used in think aloud. The only difference was that, in order to make room on the manuscript for the coding & annotation boxes, the passage was shown in a slightly narrower column, such that only one deletion *could* appear in a single line. This is discussed further below. As well as re-using the stimulus passage, it seemed sensible to try to interpret AC informants'

operations in terms of the set of categories used to analyze think aloud data. If changes later had to be made, this ought to be done on the basis of concrete AC informant data.

Some key features of the think aloud orientation sessions were taken into the AC counterpart. Firstly, explicit exemplification of coding (in the sense of specifying *a priori* how to code a given series of events) was avoided lest informants shape their own processing accordingly. Instead, potential informants were given a three step introduction to the AC task. First of all, potential informants were introduced to the codeset, and each of its components was illustrated. This presentation took place in the informant's L1 (with written backup of oral explanation) and in the TL, such that all key information was given in both languages. Informants then listened to a brief audio-recording of a JL1 helper thinking aloud in a free mixture of English and in Japanese as she processed part of the short cloze passage which they had before them. This passage had been lightly edited in an attempt to make the 'likely' clues and operations involved fairly clear, and the recording itself was edited in such a way that the speaker's mentions of the code she planned to enter were obscured by a speech-like but unintelligible masking noise. Working in pairs (Experience had shown that having Japanese L1 students work in pairs or small groups on unfamiliar tasks appeared to give them more confidence; pair members were not required to agree on common codings, however.) informants were then invited to listen again and enter the codes they felt the speaker might have used at each deletion.

(Extract from AC orientation example; L1 verbalization in small capitals)

Perhaps the best known are the (1) _____ rays or torpedoes, of which several (2) _____ live in warm seas. They possess electrical organs (3) _____ each side of the head, behind (4)_____ eyes [...]	"AH...DON'T KNOW THIS ONE...'RAY' IS A KIND OF FISH?..ANYWAY.. THE JAPANESE NAME IS...*EI*, RIGHT? ..NO GOOD..YOU NEED TO HAVE THIS.. SPECIAL KNOWLEDGE .. SO [MASKING NOISE] NEXT [...] okay two is *kinds* [...] BECAUSE I KNOW THE EXPRESSION 'several kinds of' something..SO I GUESS THAT'S [MASKING NOISE]... OR MAYBE [MASKING NOISE] WOULD BE OKAY..AH [...] NEXT ONE IS *ON* I GUESS..IT HAS TO BE A prep..a preposition SO I'LL WRITE [MASKING NOISE].. the head behind their eyes..*their* eyes? *The* eyes? MAYBE *the eyes* IS BETTER..SO IT'S [MASKING NOISE] BECAUSE.. you need to have an article between the preposition and the noun..WELL IN THIS SENTENCE ANYWAY [...]

Informants were then shown (on a 'sequenced' transparency) the speaker's actual self-codings one by one, along with (at suitable intervals) projections of the few annotations and mark-ups on her manuscript. These were discussed within the group, but as participants were not pressed to reveal their own codings it is not possible to say how closely these matched those of the speaker. My impression, however, was that for most items participants appeared to have selected the same codes as the speaker: KOW, or in one or two cases an alternatively-labelled equivalent, had been chosen for (1), 'Phr', 'Col', or 'Idi' (the speaker herself had used the label 'idiom') for (2), 'WC' for (3), and ' Gra' for (4). Encouragingly, some orientees had added the appropriate coding ' Gra' to (3), even though the speaker herself coded only her (partial) recovery in terms of the word class of the missing item. By and large, then, orientees' suggested codings were felt to bet consistent with those provided by the speaker.

Orientees then carried out the same short cloze tasks (SPILLWATCH (part) or FUNCHAL) used in some verbal report orientation sessions, noting on their manuscripts the codings that represented the processing operations they had used.

Informants were asked to process at least three to five deletions before consulting or discussing with a neighbour, and in the event most completed most or all of the task before doing so. Once again, no pressure was put on informants to reveal their success rate, but none (and I believe I have enough experience of Japanese students' reactions to have spotted even fairly mild anxiety or disappointment) appeared uneasy about her performance compared to that of her neighbour. No orientee approached me at the end of this session to 'check' her codings. This is a common behaviour among JL1 learners who fear they may have done badly in a task, but who lack the confidence to make this known during it: the fact that none sought advice seemed solid evidence that the orientees were more or less at ease with their coding performance.

7.10 Effect of passage format

I had developed a range of different task passage presentation formats and had realised that more space became available for written comments (i.e. boxes could be spaced further apart and thus effectively enlarged) if a 'column' passage format (CF) was used. This, however, extended the physical length of the CF manuscript sheet to 48cm (for the full OLYMPICS passage; the 32-deletion version was slightly shorter) , as opposed to circa 30cm for the 'regular' version (RF). This made for a rather unwieldy manuscript, so that it seemed advisable to investigate any effect these layouts might have on passage processing. I did this by observing four solo condition informants as they worked with either the RF or CF version of the passage (2 RF, 2 CF), and also two pairs of informants (1RF; 1CF) The results

were much as anticipated, though still useful. One overt difference between the processing behaviours of the CF and RF solo informants was that, perhaps unsurprisingly, the former displayed considerably more physical movement during processing than the latter. Compared to RF passages, the greater physical length of the CF passage did not appear to lead to longer 'search' times (though it is not easy to tell with any certainty or precision when a search begins or ends) although on a few occasions informants were observed to momentarily act as though they had 'lost the place' in the passage. This occurred following their insertion of codes on the separate coding sheet used in this trial, and was one factor behind my decision to incorporate everything onto a single manuscript.

I had anticipated that the CF condition might prove irritating to informants precisely because of the greater length of the task sheet. On being shown, post task, the format (CF or RF) which they had *not* used, informants in fact commented first on the comparative lengths: "AH! PRETTY LONG!"[CF] ; "WHY IS IT SO SHORT? OH, I SEE. IT'S WIDER." [RF] and on the relative usability of the two formats: "MAYBE EASIER TO WORK WITH?" [RF].

On the other hand, two solo informants who had completed the CF task before seeing the RF version commented that, while the CF version was cumbersome, it made it possible to locate deletions more easily as there was only one per line. In only a very few instances did RF condition informants appear to need more space for their comments than was available, but I wanted to elicit informant comments on the task as a whole. To this end, all solo informants were shown (during a brief

post-task question stage) the 'other' version of the task, and given this prompt:

> *"I have these two version of this passage, and I need some help in choosing which one to use. Looking at them both, does either one seem better?"*

The consensus of solo informant opinion was that the RF was probably more manageable, but in terms of how it would affect processing they could see no important differences between the two formats. None of the solo condition informants expressed a very marked preference for either.

'Navigation' by paired informants

Just as I had done with think aloud, and again in order to find out whether the conditions affected how much was reported (or, in the case of AC, how many codings were applied) I had some AC informants work in pairs. Paired informants had more obvious 'physical' difficulties with the CF task: on several occasions I noted one partner moving the sheet or herself in a way that seemed to interfere with the other's processing. I was already aware from paired think aloud that few pairs work entirely in synch, as it were, and it seemed reasonably clear that the need imposed by the CF task condition to 'navigate' greater distances was itself creating additional difficulty for the informants. To minimise the problem of 'incompatible navigation' within the passage, I resolved to keep the entire passage on a sheet roughly the size of two B4 pages side by side, formatted as shown in Figure 7.6, below. Previous AC informants to whom it was shown agreed that this format was more manageable than the CF format. The format shown below was able to accommodate the full, forty-one deletion OLYMPICS passage. When this

was shortened (see below) to thirty-two items, it became possible to use a single-column presentation in a vertical layout.

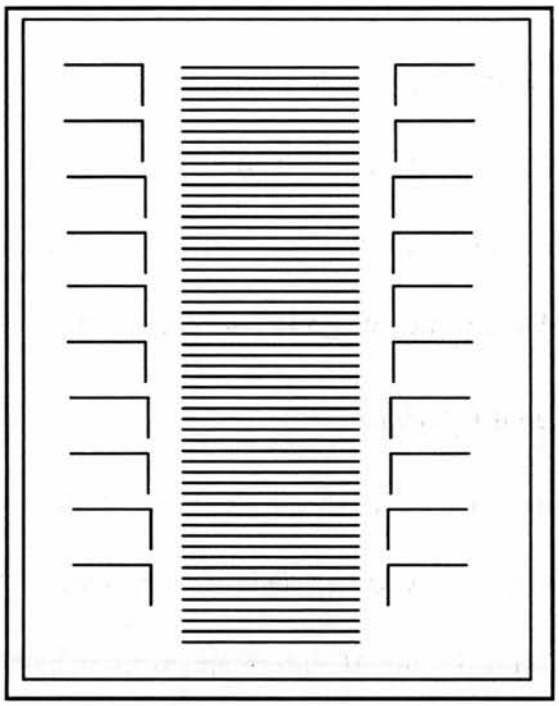

Figure7.6: Final AC manuscript format, with numbered boxes flanking passage.

A good deal could be said here, if space allowed, of my experience of trialling pair- (PCAC) vs. solo-condition AC (SCAC). I will confine myself to explaining the three main reasons why I felt it better to abandon the former in favour of the latter. The first reason was that, due to increased demands on students' time, the number who could make themselves available as informants decreased quite dramatically. Secondly, and just as importantly, it appeared from the trial sessions conducted that PCAC offered little in the way of additional insight that could not be gleaned from SCAC sessions. Much as in pair-condition think aloud, a proportion of task time was taken up with phatic and other only tangentially task-related interactions between partners (although these no longer had to be

audited and transcribed.) Last but not least, among those potential AC informants to whom I had introduced the task format, there did not appear to be anything like as strong a preference for pair-condition reporting as had been noted in the think aloud groups. This was unexpected, and suggested that some of my explorations of PCAC had been for nothing. I retained the task sheet presentation trialled with paired informants, however, as this had proved equally satisfactory in solo-condition. I then turned my attention anew to the question of how best to present the codeset to informants.

7.11 Acquainting informants with the codeset

The question of how AC codes could be more efficiently made familiar to informants in small-group orientations was a pressing one. The mode of presentation in use at that time was essentially a simple list, in which codes were grouped according to their physical 'closeness' to the deletion (syntactic relations, phrase-element status, collocational links, etc.) or by some other criterion, such as the 'route to recovery' superordinate covering, say, knowledge-of-the-world, logical relations, etc. This 'list' approach was far from unusable, but informants seemed to spend more time than I would have wished peering at the list of codes. At some point I realised that it might be far easier for informants to become familiar with codes if these were presented in a kind of 'node and branch' (a.k.a 'decision tree') format, such that making one selection led on to a manageably limited set of further options.

I thus began to work up possible hierarchical arrangements of the codeset, only to note fairly quickly that even from my own standpoint it was difficult to choose between alternative arrangements which seemed to make equal sense: should, for example, WC (denoting the informant's delimitation of a filler to the class of verbs, nouns, etc.) derive from the 'Gra' node, or did it rather belong with other codes denoting 'partial success' such as 'SWA' or 'Und'? I had placed 'WC' under 'grammatical' based on my own intuition that the realization that an unknown filler must be a verb reflected primarily an awareness of grammatical relationships. Others, however, saw it as rather pertaining to vocabulary. On consulting a number English and Japanese native-speaker colleagues, it became clear that different arrangements made sense to different people, and that it would be extremely difficult to construct a single arrangement which embraced all the codes in the set. The value of a presentation system that allowed me to group codes according to 'common features' seemed nevertheless real enough, and this notion was confirmed when I invited a small 'focus group' of former think aloud informants (who were thus already familiar to some extent with the idea of describing processing behaviours and thus to the need to label and analyse these) to comment on the codes themselves and on the possible arrangements I had drawn up:

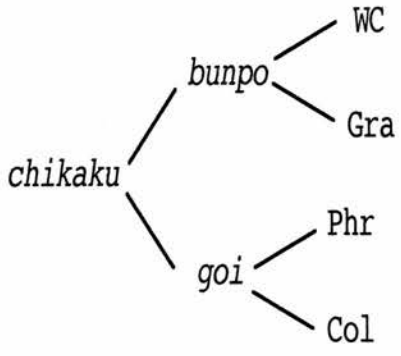

Figure 7.7: 'branch' presentation of codings of local context cues

The figure above shows my own arrangement of the sub-set of 'chunk' codes, all of which pertained to local-context cues, and in which only the location of 'WC' was seriously questioned by consultants. It seemed sensible, too, to label nodes in the informants' L1 as shown here: '*chikaku*' ('near'), '*bunpo*' ('grammar-related') and '*goi*' ('lexical'.) The extent of the differences in the relationships among codings perceived by different people, however, left me concerned that in setting up my own node-and-branch system I might end up confusing as many informants as I helped. The second (and if conceptually trivial, in practical terms quite serious) problem with the node-and-branch presentation was that of printing up legible copies. In order to present with any clarity the codeset and even abbreviated definitions in this format, at the very least an A3 page was required. Using this size of page, plus a manuscript, on the small desks common in Japanese classrooms meant that either the codeset had to be under the manuscript, or vice versa. In an LL booth (and I wanted to keep open the option of having AC informants think aloud at the same time) it was impossible to use a sheet of this size.

Not surprisingly, the end result was a compromise, but one which I think represents a worthwhile step forward from the 'list' and 'node-and-branch' options. Obviously, someone had to prepare and co-ordinate revisions of the codeset, and that someone had to be me. I thus prepared the list of codes shown in Appendix 2, with codes grouped together in ways that seemed reasonable to me, and indeed to some of those I consulted. Actual presentation to informants of the items on the list, however, changed radically. Orientations became much more interactive sessions, in which 'codeable' behaviours were first elicited from informants via questions: "What would you do after you turn over your test page and see the passage?"; "What could you do if you're stuck and can't find a word?" and presented on B5 cards as they were volunteered, with first a paraphrase of what had been proposed, then the coding on the reverse:

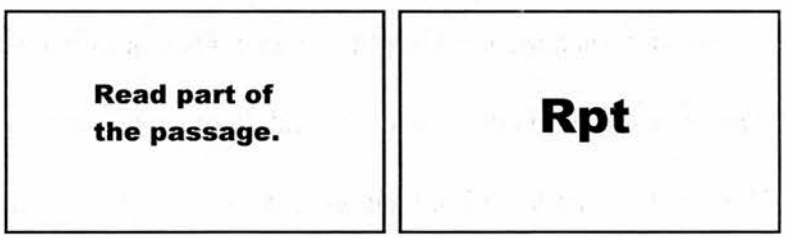

Figure 7.8: Card showing individual coding and rubric

In this way, what one informant had described as a "scary" list was assembled in stages. Informants were then put in pairs, provided with a copy of the printed codeset and a set of smaller business card-size versions of the cards used in elicitation of behaviours, and invited to group these in the way(s) that seemed most plausible to themselves. Two blank cards were also provided to each informant so that she could (none did, however) add any codes of her own. Those

who wished to do so were also encouraged to draw up their own layout of codings for use as an aide-memoir or reference (Nevo 1989 notes that checklist users need to have the list to hand during the task for successful application of the codes the list contains.) for their individualised (sub)set of the full codeset, which might then be very much secondary if not superfluous.

This use of the arrange-it-yourself card set also appeared to help informants to think about which behaviours were more central to their own processing and which might be peripheral—in a sense giving them permission (and this, I think, is a critical feature) to set aside any that seemed irrelevant to their own needs. It was fairly clear from observation that orientees would typically work up a smaller set of codings which–questioning confirmed–they felt most applicable to their own processing behaviours. While orientations did become longer, there is no doubt in my mind that they also became much more effective and perhaps almost enjoyable for participants. I noted, in the products of these sessions, fewer 'odd' structurings of codes than I had anticipated. Rather than attempt to make the more idiosyncratic arrangements conform ('Conform to what?' was the obvious question.) I limited my intervention to making sure that the informant had a clear idea of what each individual coding meant to her. If that understanding was present, and communicable to myself, I could see no real grounds for querying the lumping together (as one informant had done) of 'Gra' and 'Und'. In section 7.15, below, I present some of the individual palettes of codings constructed by AC informants.

7.12 Some limitations of AC in tracking processing events

It must be admitted that potential shortcomings in AC procedure were apparent. One point which was raised by a JL1 consultant (who had herself encountered cloze tasks in ESL classes prior to and during undergraduate studies in the USA) was that the AC format might lead informants to focus on and record their processing, as it were, blank-by-blank (cf. the issue of 'co-processed' deletions in NCR in chapter 6) with the danger that 'wider' behaviours such as the linking of deletions might get lost. This caveat was echoed in the observation of a former pair-condition think aloud informant to whom I explained the AC task in its then current stage of evolution, and asked for his comments.

This informant pointed out (and here I paraphrase extensively) that the AC task as it stood did not prompt informants to mention any problems they were having in comprehending the extant passage, i.e. the unmutilated content. She could. She claimed, clearly recall that she and her partner had discussed the meanings of several extant words and phrases from their OLYMPICS cloze passage, even she could not recall what most of these had been. I had somehow managed to overlook this missing element in AC up to this point, even though think aloud protocols had provided quite clear evidence (if such were needed) to support the suggestion (cf. Alderson 2000) that the challenge of cloze need not reside solely in the deletions. It was, however, not easy to see how to get at events 'lost' in AC except via post-task questions—something the procedure was intended to do as far as possible without. As for the tracking of difficulty in comprehending extant

passage items, I hoped that this would be recorded via the coding 'UPI' (denoting an 'unfamiliar passage item') and/or via the recording via underlining of translation or glossing into the L1. In the event, I added an oral instruction in orientations, encouraging informants to "feel free to note down" (*jiyu ni kaite kudasai*) any observations about their processing which they felt the AC manuscript did not adequately record, or anything they wished to tell me about their reactions to the task as a whole. These comments could be written in the L1 on a separate sheet of paper, and if anonymity was wanted this did not have to be attached to the informants' manuscripts.

Very few comments were obtained which pointed up shortcomings informants perceived in the AC procedure itself, besides suggestions that the passage was "very difficult" in whole or in part, that informants be allowed to compare ideas (apparently post-task), or that a strict time limit be set for the task. (Some informants were seen to complete the task some time before they formally handed over their manuscripts, and it may be that they had been reluctant to finish conspicuously ahead of others.) No informants offered other negative comments in the short PTQs or 'debriefings' I carried out, and only a very few informants added comments on a separate sheet of paper. These also included some of the above, but for the most part took the form of L1 or TL apologies-in-advance for poor performance: "I could only write some answers"; "There may be many mistakes", etc. These might also be seen as further evidence (cf. chapter 5) for my suggestion that Japanese L1 informants find it difficult to treat 'test-like' tasks

very differently from actual tests.

One further potential problem with the AC task as used here also relates to the item-by-item aspect mentioned above, but was only noted after much of the data had been gathered. The problem has to with where the informant should the informant insert codings for initial steps, if used, such as pre-reading of the passage. A very few informants did (see below) insert the coding 'Rall', indicating that they had read over the entire passage, and two of these applied the coding to the first blank. While this was the best place available to them, it is arguably not completely appropriate, and there is a possibility that more informants read the passage before beginning to fill blanks, but did not record this for want of an appropriate place to do so. A checklist-based task such as AC needs, then, to make provision for events or behaviours not directly linked to a given deletion–although a possible counter argument would be that a location on the manuscript in which to note or code for pre-recovery behaviours could easily lead informants to assume that these were expected, and modify their 'natural' cloze processing to suit. The 'enveloped' questions outlined above may be a viable, if cumbersome, solution to this point.

7.13 *ab initio* categorization & labeling of events by informants

In a limited sense, the 'interactive' orientation session outlined above, with its elicitation of intuited processing events, can also be seen a validation exercise. Time in orientations, however, did not allow for extensive informant discussion of codes as they 'emerged', and there was no disguising the fact that I had a set of

codes ready to produce one-by-one. In the hope of further validating the codeset against the intuitions of informants, I arranged, simultaneously with the ongoing orientation sessions, the small (four person) focus group exercise outlined below. These informants were recruited *ad hoc*, and although all had at least some experience of didactic cloze or gap-filling tasks, none had previously served as an informant, or been oriented to the AC task. Initial questions indicated that the four, all women, had no real knowledge of the AC task at all, beyond an awareness that I was looking for informants to take "a kind of test." The four were shown the 11-item REFUGEES passage, above, and the nature of the cloze task explained just in case some could not recall how to carry it out. Participants were then asked to fill the first few blanks and then to talk about how they had done so.

I was relieved to find that the observations respondents made about their own processing of the cloze blanks could in every case be mapped without much difficulty onto an existing category or event, although the limited demands of the passage meant that the set was not fully exploited. Although the conversational nature of a group task does not allow an auditor to know precisely how many individuals in the group actually did *x* or y, I was able to satisfy myself that these respondents had used knowledge of phrases or collocations in, for example, items (1) and (4), and knowledge of grammar in (3). They had also clearly accessed information earlier in the paragraph ('SPb') in filling items (5) and (7), and extratextual knowledge in (9). Uptake of a cue ('causes') from an earlier paragraph ('EP') had also been used in deriving the incorrect filler 'cause' in (11).

My intention had been to test informants' 'naïve' descriptions of their processing behaviour against a wider set of coding categories, but time did not allow me to set a second and longer cloze task. Instead I distributed copies of the OLYMPICS passage and asked respondents to identify blanks which could be filled by the same methods they had just used. The group consensus did not differ much from what had already been seen in think aloud and NCR protocols, and which would later be found in AC data: items such as (1) required "special knowledge"; (3) could be filled by reference to previous and following sentence information, etc. As a final step, I read aloud to the group some of the definitions given on the reverse of the codeset cards, and asked them to match these to appropriate blanks. Again, this exercise produced no real surprises, although the speed with which matches showed considerable variation. Unsurprisingly in view of the minimal cue-target item distance, the most quickly coded deletions were those recoverable by access to grammatical ('Gra') and phrasal/collocational ('Phr/Col/Idi') knowledge. (One previously, and indeed subsequently, un-met event was the application by one informant in the group of 'LP' (information from a later paragraph) to deletion (7), by which she explained her non-SEMAC filler of 'official victors'.)

To sum up, conflicting pictures may be found in the literature of how well informants can match their behaviours to a checklist (Nevo 1989; Allan 1995) and this brief session offered some reassurance that informants could in fact identify or describe their own processing behaviours, and also match these to the

categories found in the existing codeset even without the deliberately limited orientation and practice I had been giving. The AC task, then, seemed to be within the capacity of informants.

7.14 'Ranking' of AC codes & individual AC repertoires

A higher proportion than expected (ca. 45%) of JL1 AC informants, including a number who in the end did not participate in data-elicitation sessions for one reason or another, attended more than one orientation session. Whatever the reason for this, informants who had attended more than one such session had undergone more practice (Had I anticipated repeated attendances, I might usefully have collected the first practice task manuscripts for comparison with those from subsequent sessions.) and as such were in a good position to comment on the relative salience of codes. Ten respondents who had repeated an orientation session (Some had also completed the AC data-elicitation task at this point.) were asked individually to convey their ideas about which codings had been more important or more useful in her own processing of passages in orientation tasks.

Rather than ask each respondent to try to 'rank' codings in order of importance—a time-consuming and difficult task—I showed her a set of twelve cards bearing the individual codings listed below, and invited her to place each one under one of three headings labeled 'very useful' (*totemo koukateki*) 'quite useful' (*ma-ma koukateki*) and 'less useful' (*amari koukatekidewanai*).The codings listed here were not selected arbitrarily, but reflected a mixture of commonly applied and less commonly applied codings about whose relative value

or applicability I hoped to gather respondents' conceptions: How would 'local' and 'distance' cues be evaluated? How much of a role would be given to prior knowledge or to translation? The percentage allocation of codes to headings by the ten respondents are shown below:

	'Very useful'	'Quite useful'	'Less useful'
Gra	100	0	0
Phr/Col Idi	90	10	0
KOW	30	50	20
SSb	50	40	10
SSf	40	30	30
SPb	40	30	30
SPf	20	40	40
EP	30	40	30
LP	20	40	40
Tr	40	40	20
Rpt	50	50	0
Rall	10	50	40

Figure 7.9: Ten informants' rankings of the applicability of 12 AC codings

One shortcoming of this session will be clear: respondents were asked to weight codings from a fixed set which may or may not have corresponded to their own palettes. The task might have been more authentic had respondents been asked to allocate only items they themselves held in their palettes (to which they did not refer during the task, although a set of cards bearing the codes and rubrics for all the codings used here was made available.) That said, I wanted to see how much weight would be given to a coding which appeared to be comparatively seldom applied in practice tasks ('LP') as well as one which I was fairly sure was not overt in any informant's palette ('Rall'.) These allocations were also 'abstract' in the sense of not being tied to a particular task, but all the respondents had previous experience of matching codings to cloze deletions, as well as of didactic cloze and

gap-filling classroom tasks; hence they did not lack concrete experience of the task.

The results, I think, are interesting. Grammatical and phrasal knowledge were obviously highly valued, while the fairly small differences in 'weight' attached to SSb and EP vs. SSf and LP (which share exactly the same distribution of weightings) may have indicated an awareness that the former were more commonly applicable. The role of extratextual knowledge ('KOW') seemed to be seen as middling (Log(ical inference)' was not listed as it hard to separate it from extratextual information; cf. Jonz's discussion of conflicitng notions of 'extratextuality' in Oller & Jonz 1994:333) while translation seemed to be accorded much the same level of importance. Reading ahead is clearly seen as valuable, but the code 'Rall', denoting reading of the entire passage, appears to be given considerably less weight. This coding was in fact rare, and the act of reading the entire passage seldom observed among informants. ('Rall' was only found on one GL1 and two JL1 AC manuscripts, or 7.5% of the total. On the GL1, and on one JL1 manuscript, the code was attached to the first deletion, and in the other it was applied to the final item, apparently indicating a *re*-reading of the full passage.) We might infer that respondents on this task slightly exaggerated their perception of the role of 'Rall', but it is equally plausible (as I suggested above) that more informants might have applied the coding had there been provision on the manuscript for separate notation of events not specific to a particular item. Interestingly, Allan 1992a found pre-reading, apparently of the full passage, to be

a very common behaviour among Chinese cloze test-takers, but while Japanese students may skim-read a conventional text in full before reading more carefully a second time (This, on the basis of a reasonably extensive survey of graduates of Japanese high schools, appears to be a widely 'taught' strategy.) there may be something about cloze (but, as noted above, less about lexical inferencing tasks, which may be the closest analogues to cloze in widespread use) that discourages this behaviour. It is tempting to see this reflected in the label of 'puzzle' by which my own JL1 students have frequently been heard to refer to didactic cloze passages–implying that these are seen as something other than 'authentic texts.' However cloze passages are perceived, the overall allocation of codes to categories, above, strongly suggests that task-takers perceive the relative importance of more local cues, and recognise the potential usefulness of prior knowledge and translation into the L1.

7.15 Individual coding palettes

Given that during orientation sessions informants' individual 'palettes' of codes were often still fairly fluid, I was interested to know how large a set the typical informant chose to work with. Owing to schedule constraints I was unable to schedule even a group interview with JL1 AC informants about their palettes of codings (and–seeAppendix 4–no interviews with GL1 informants were possible on this point due to time constraints) I asked informants–all or almost all of whom had by this time already completed the OLYMPICS data-elicitation task–to circle on a preprinted list of codings those which were currently in theπir own palettes.

data has serious shortcomings, but I report it here for the picture it gives of individual coding sets. Eleven informants complied, and their palettes are shown below in descending order of size. 'Phr(rase)', 'Col(location)' and 'Idi(om)' are grouped here as discussed in chapter 6, so that the circling of any of these on the form is represented by the conjoint coding 'Phr/Col/Idi', below. Due to an oversight on my part 'SS' (same sentence) and 'SP' (same paragraph) were not subdivided on the form, with the result that up to two finer codings from an informant's palette ('SSb'; 'SSf') may be missing. Other limitations are discussed below:

1 Gra; Phr/Col/Idi; KOW; Log; Tr; Rpt; SS; SP; EP; LP; LFN; Sou; GS (13 items)
2 Gra; Phr/Col/Idi; KOW; Tr; Rpt; SS; SP; EP; LP; LFN; Sou; GS (12 items)
3 Gra; Phr/Col/Idi; KOW; Log; Tr; Rpt; SS; SP; EP; LFN; Sou (11items)
4 Gra; Phr/Col/Idi; KOW; Log; Tr; Rpt; SS; EP; LFN; Sou (10 items)
5 Gra; Phr/Col/Idi; KOW; Tr; Rpt; SS; EP; Sou; LK (9 items)
6 Gra; Phr/Col/Idi; KOW; Tr; Rpt; SS; EP, GS (8 items)
7 Gra; Phr/Col/Idi; KOW, Tr; SS; EP; LFN; Sou (8 items)
8 Gra; Phr/Col/Idi; KOW; Tr; SS; LFN; Sou (7 items)
9 Gra; Phr/Col/Idi; KOW; SS; EP; GS (6 items)
10 Gra; Phr/Col/Idi; KOW; Tr; Rpt; SS (6 items)
11 Gra; Phr/Col/Idi; KOW; Tr; SS; GS (6 items)

Figure 7.10: Individual 'palettes' of codings held by JL1 AC informants.

The broadest palette submitted here, then, has 13 entries, the narrowest 6, and the mean is 8.7. From my recollections of observations of and discussions with informants during orientations, 13 codings is at, or close to, the high end of the range, while six may be at the low end. Owing to another oversight, the preprinted

form had a space for the informant's name but no instruction to enter this. Only four of the 11 informants who returned forms provided their names, and this is clearly too small a sample to associate breadth of palettes or type of entries with task success as measured by SEMAC scores. AC informants were not allowed to take away copies of the tasks they completed–even practice tasks–and as it is unlikely that any had a cloze passage or two handy on which to base their circling of codings it is hard to imagine that respondents would have amended their palettes (which they may or may not have consulted) especially for this elicitation. Indeed, the fact that three contained only six entries suggests an honest report on the form provided. It is, however, possible that informants retained in their palettes codings which they anticipated might be useful, but which in practioce they seldom applied (cf. the closer-than-expected weightings of 'EP' and 'LP' in Figure 7.9 above.) This aspect of the AC task may merit further study, but this could be achieved only through sustained on-task observation and face-to-face interview.

At 46% of the total, and apparently representing the full range of 'breadth' of palettes, there is no reason to think that those shown above are unrepresentative. Making the assumption that this is a representative sample, we can see that few (here, only two) palettes contain the coding ('LP') for information in a paragraph subsequent to that containing the target blank, whereas eight contain that pertaining to information from an earlier paragraph ('EP'). This also seems to be at odds with the data in Figure 7.9, suggesting that the importance of a given code

may differ in more, and less, abstract contexts. Surprisingly, too, considering the weightings 'SPb' and 'SPf' in the same figure, only three palettes contain the coding for information outwith the sentence containing the blank, but within the same paragraph. The coding 'Sou', representing evaluation of candidate fillers by how they sound in context appears in seven palettes, with that denoting evaluation by appearance ('LK') alongside the former in one set. This corresponds with observed behaviour in verbal reporting, in which evaluation by sound appears to be more common, and to a lesser extent (as evaluation events are less reliably coded for) in AC itself. The code for recovery via 'guessing' appears in five palettes, which is surprising insofar as the coding seems not to be widely applied in AC, or indeed widely applicable in verbal report. I would speculate that it represents a kind of 'fallback' coding which informants wish to keep in reserve for situations where no other coding can be applied. Comments from two AC informants seen to hold this coding in their palettes suggested that something along these lines was indeed a factor: neither informant could recall actually having applied the coding to an item, but both claimed to "sometimes guess."

Another point of note is that six palettes contain the coding 'LFN', indicating a temporary abandonment of the blank. This is another seldom-applied coding in AC, and on a number of occasions I have observed informants quite clearly leave an item without coding for the behaviour. Post-task questions on this point elicited from one informant the remark that he expected to "quickly" find information that would help him fill the blank, and it may be that it did not seem worthwhile

entering the coding 'LFN' for what was expected to be a brief interval. Think aloud data suggests that abandoning an item is something of a last choice, and it is quite possible that verbal reporters only explicitly do so when they have given up on finding cues in the surrounding text: it is not hard to imagine that, given responsibility for coding their own processing events, AC informants (perhaps unconsciously) avoid or delay signaling lack of success. These perhaps minor anomalies aside, the palettes listed above make clear their owners' awareness of the central role of immediate cotextual cues to recovery as denoted by 'Gra' and 'Phr/Col/Idi' and 'same sentence' local cues. The role of extratextual knowledge also seems to be well-established, as is that of translation and reading ahead—even though these last two codings are not very commonly applied by AC informants.

To sum up, the individual palettes above seem to raise as many questions as they answer. Do AC informants select only codings they actually apply, or are codings such as 'G(ue)S(s)' and 'L(ater) P(aragraph)' included as hypothetically useful rather than much used in practice? (Even if AC informants' palettes of codes are not identical with the codings they actually apply, the former very likely constrain the latter.) Could it be that the weightings in Figure 7.9 actually better represent the role of events such as translation ('Tr'), and that these are under-indicated in AC? To answer these questions it may be necessary to 'follow' AC informants over a longer period as they construct their coding palettes and revise these in the light of interaction with cloze passages. Such a 'case study' approach may prove

fruitful, but demands a perhaps unrealistically high level of commitment from those being studied. A second-best alternative might be to ask informants to verbally report on their actions as they carry out the AC task, a small and only moderately insightful trial of which procedure is reported below.

7.16 AC under verbal report conditions

In an attempt to gain more insight into the strengths and weaknesses of AC, and in particular into what information might go unrecorded in the format, I asked five JL1 informants without prior experience of the task, or orientations to it, to carry out an abbreviated version of the OLYMPICS cloze task under a mixture of AC and think aloud conditions (The target passage, cut down to fit the time slot available, incorporated deletions (5) to (21) with passage content prior to the first blank left unmutilated.) i.e. to code their processing behaviours while simultaneously verbally reporting what they were doing. The orientation to both methods was necessarily shorter than I would have liked (although I believe it was adequate) and more practice might have helped generate more sustained verbalization in the actual task. The verbal data that emerged from this session corresponded more to NCR-style immediate retrospection than to continuous thinking aloud. Some excerpts from protocols will illustrate this (L1 verbalizations are shown in small capitals. with selected AC coding in brackets):

Deletion 21/Informant 1

"[...] I DON'T REALLY KNOW MUCH ABOUT THIS pentathlon.. I THINK IT MAY BE jumping BECAUSE THAT'S AN OLYMPIC SPORT NOW.. MAYBE IN THOSE DAYS IT WAS THE SAME {...] [KOW]

Deletion 5/Informant 2

"[...] NUMBER FIVE.. no one SO I'LL CHOOSE PHRASE [...]" [Phr]

Deletion 7/Informant 3

"[...] came from all parts of Greece? SO EASY?.. IT'S A SPECIAL PHRASE [...]" [Phr]

Deletion 14/Informant 4

"[...] FOURTEEN.. exact AND [connotation of goes with'] number.. MANY SPORTS ARE MENTIONED SO IT MAY BE OKAY.. phrase [...]" [Phr]

Deletion 19/Informant 5

"[...] THIS IS DIFFICULT... MAYBE varied IS A VERB SO I PUT that .. THEN AGAIN.. CAN'T [DO?] MORE.. SO GRAMMAR RIGHT? [...] [GRA]

Post-task questions (1): insight added through verbal report

During the session the informants did not verbalise in much depth, and it was clear that their attention was primarily on the AC task. The informants were asked, almost immediately post-task, to individually review three deletions apiece. The deletions were allocated to each informant not in a random fashion, but (as the chart below shows) to cover almost all of the items included in the task. No account was taken of whether each informant had or been heard to verbalise about the deletions she was now asked to consider. Two questions were put to informants, with L1 written translation. Firstly, I asked:

> *"Listening to your recording again, how much information do you think it adds to the information on your manuscript about how you dealt with blanks X, Y and Z?"*

Once I felt sure that the question had been clearly understood, I asked consultant-informants to listen on headphones to their think aloud and to choose a value between 1 and 4 on a scale (set out in the TL and in the L1), which represented 1 'nothing really' (*hotondo nai*), 2 'just a little' (*sukoshi aru*), 3 'a useful amount' (*kanari aru*) and 4 'a lot' (*takusan aru.*) While or after listening to

as much of their protocol as they wished to (winding on was allowed) each informant either selected a numerical value directly or echoed the scalar representation of what it represented, e.g."just a little". The distribution of target deletions is shown below, with informants numbered down the left side of Figure 7.11:

Deletions targeted

	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1				X							Y						Z
2	X				Y								Z				
3			X			Y			Z								
4		X								Y				Z			
5							X					Y			Z		

Figure 7.11: Allocations of AC / verbal report items for review.

	Del X	Del Y	Del Z
Informant 1	1	2	3
Informant 2	1	1	1
Informant 3	2	1	1
Informant 4	1	3	2
Informant 5	2	1	2

Figure 7.12: Informants' ratings of informational content added through verbal report to that of AC.

As table 7.12 shows, only two informants chose a value above '2', and '1' represents just over 50% of the values chosen. This suggests that, overall, verbal report was not perceived by informants themselves as adding much insight over what they had recorded via AC. When we look at the deletions for which verbal

report does seem (to informants themselves at least) to have 'added value' we see that these are (21), and (14), both deletions predicted to be at least moderately difficult by GL1 and JL1 consultants, and filled correctly (SEMAC) by 62.5% of GL1 and 56.2% of JL1 AC informants. One conclusion I draw from this data is that verbal report, in its much more labour-intensive think aloud form, might most efficiently be used to target items previously identified by AC + verbal report as more open to added insight..

Provided informants can be orientated to the reporting formats involved, then, there may be benefits to combining introspective data-elicitation task formats. Items which AC shows to be consistently cued by immediate-context grammatical information, for example, may be less worthy of think aloud attention (which, as noted above, may in fact not even reliably record activation of these cues) than items which appear to be recoverable in more than one way, or which are open to a range of interpretations. In such cases the return on transcription time, etc. may appear more worthwhile. If, moreover, (as I suggested in chapter 5) the focus of didactic application of cloze-type tasks shifts from fixed-ratio cloze to gap-filling/rational cloze, think aloud reporting may be more consistently productive and worthwhile.

Post-task questions (2): changes to written comments

The second question I put to the five informants was:

"Thinking now about showing how you [filled the blank], would you add, take-away or change anything you wrote down [in the space provided for comments]?"

Through this question I hoped to get some insight into informants' perceptions of the value of any written comments they had added to their AC manuscripts. Such written comments had been fairly sparse, perhaps because informants felt writing to be superfluous when a microphone was at hand. Informant 1 suggested, I think seriously, that she would add a comment that deletion (21) was unfair as it required special knowledge of sports. Interestingly, one deletion had attracted a similar comment from four of the five informants: in deletion (21) (SEMAC fillable by any of a range of sports events) all four noted unfamiliarity with or uncertainty about which events comprised the pentathlon, and one or two appeared to mistake it for the triathlon, a better-known and more fashionable sport in Japan today. Other written comments offered alternative or 'second choice' fillers for a variety of blanks. Informants were fairly open about the value of, and motivation for, adding alternative fillers. Two used the L1 expresssion '*daiji o toru*' ('play it safe') in relation to their inclusion of alternative fillers, suggesting they saw the opportunity to offer a second-guess as a kind of insurance or 'face-saving' step. Perhaps the most tantalizing suggestion came from the informant who seemed to imply that being able to note alternative fillers had allowed her to, as it were, move on instead of struggling to choose the best one. It might be interesting to see how the option of inserting more than one (assumed to be) SEMAC filler in regular cloze blanks would affect the time test-takers spend on task, but that is another study.

To sum up, we have seen that these AC informants themselves seemed to feel that their verbal reports added only moderate insight into their conduct of the AC task, suggesting that they had either found the task too burdensome or did not think much could be added through verbal report. Looking just at insights rated as '3' by informants, one may wonder how much useful insight Informant 1's verbal output added to his filler and coding ('jumping' and 'KOW') added in deletion (21). In deletion (14),on the other hand, the verbal report data suggests that informant 4 was in fact using not just the phrasal/collocational association of 'exact' and 'number' that his AC coding of 'Phra' suggests, but that he had also taken into account the 'SSf' and 'LP' listing of sports events. This seems to be information worth having, and suggests that AC may tend to record only the (to the informant) most obvious or salient coding.

I might add here that, in regard to the post task questions above, I was criticised by a JL1 consultant for putting informants in what she saw as an uncomfortable situation–one in which their AC+verbal report performance was implicitly deemed inadequate. I do not think the post-task review was interpreted that way by the informants, but I may have been insufficiently sensitive to the possibility. Finally, to answer the kinds of question posed in section 7.15, focusing on whether AC 'loses' or under-represents certain events (as it fairly clearly does) and which ones, a larger and better designed study than the exploration reported here would be needed. This might target only those deletions which (as suggested above) seem to be more amenable to thinking aloud *about.* This, combined with

more extensive orientation and practice than was possible here, might reduce the added burden of verbal reporting to a level at which informants would be able to generate more insight about their interaction with the passage and their filling of blanks.

7.17 How 'stable' are the codings applied in AC?

Having established to my satisfaction that informants could describe and match-to-categories their processing behaviours, I thought it desirable to look at whether AC informants selected similar AC responses to similar items in different passage contexts. This was something I had hoped to do with think aloud informants, but found at the time that I lacked an adequate number of informants, sufficient time to analyze the products, and sufficient access to recording facilities. There was, however, no real barrier to my at least attempting this step with AC. As there was no overlap between participants in the 'pure' think aloud data-gathering rounds and those taking part in this AC round, there was also no barrier to the recycling of two passages (REFUGEES and SPILLWATCH, see appendices) which had already used in some of the think aloud practice sessions, in a kind of test-retest procedure. Being entirely paper-based, moreover, this AC task could be carried out in lesson time to plausibly pedagogical ends.

A 'seminar' class of eight junior year students (all intending teachers and thus with some interest in classroom research) was used for this small study. The codeset was introduced as outlined above, and informants were oriented in its use by having them apply and discuss codes on sample cloze passages. The following

week, each informant individually processed one passage in AC format, half (n=4) receiving REFUGEES and half SPILLWATCH. After an interval of 14 days, those who had received REFUGEES first were given SPILLWATCH (Order A), and those who had received SPILLWATCH were given REFUGEES (Order B.) Because of one informant's absence from a session, sample sizes were matched by selecting for comparison three informants from each order who had completed both passages. Mean SEMAC scores across both passages for Orders A and B (n=6) were compared via independent samples t-test, and no significant difference reported. Passage difficulty, then, appeared to be comparable. The 'stability' with which informants coded AC deletions was checked via examination of their codings for five deletions in the REFUGEES and SPILLWATCH passages which were felt to be directly comparable, namely items (3); (7): (14); (21), and (22.) The results are shown in the table below, with each informant's chosen filler shown in italics below the coding(s) she applied to that deletion.

The idea of the two parallel passages REFUGEES and SPILLWATCH was that some of their 'paired' deletion contexts should elicit the same processing behaviours. As the table shows, informants have applied at least one of the same codings in 26 out of 30 (86%) instances, though in not all cases was a SEMAC filler supplied. (Informants may of course know which cues point to a filler without being able to supply the filler itself, or vice versa, so that this last point is not of central concern here.)

Blank	SCAC1 (A)	SCAC2 (A)	SCAC3 (A)	SCAC4 (B)	SCAC5 (B)	SCAC6 (B)	Remarks
3 Refug	Phr to	Gra to	Gra to	Phr to	Gra to	Gra to	
3 Spill	Phr to	Gra to	-- to	Phr to	Gra to	Gra to	SCAC3 omits coding
7 Refug	SPb Today	-- --	SPb Nowadays	SPb Today	-- Now	SPb Now	SCAC2 omits coding, filler
7 Spill	SPb Today	SPp Now	SPb Now	SPb Today	SPb Now	SPb Today	
14 Refug	Gra that	Gra that	Gra that	Gra that	Gra that	Gra that	
14 Spill	Gra that	Gra that	Gra that	Gra that	Gra that	Gra that	
21 Refug	Phr capable	Phr capable	Col capable	Phr capable	Phr --	Col capable	SCAC5 omits filler
21 Spill	Phr of	Phr, Col of	Col of	Phr of	Phr of	Col of	
22 Refug	SPb displaced	SPb, LAN displaced	LAN --	Phr --	SPb displaced	SPb, Phr displaced	SCAC3,4 omit filler
22 Spill	SPb agree-ing	SPb, LAN agreed	SPb, LAN agreed	Phr agree-able	SPb agreed	SPb agreed	

Figure 7.13: Codings of comparable items across two cloze passages

Three arguments might be raised against the attempted test of stability above. Firstly, the paired passage items may simply be *too* parallel: adjective +infinitive in deletion (3), find+that in (7) and the same phrase 'capable of' in (21). Secondly, the paired contexts all elicit fairly local cues: Gra, SPb, Phr, etc. These would be fair criticisms, and more sophisticated parallel passages might usefully be constructed to assess whether longer-range cues are also applied with similar

stability even when the parallel structures have less in common. Thirdly, the fact that three out of six informants filled deletion (7) with exactly the same word in both passages (and one other used a shorter variant) suggests a memory effect at work across the passage administrations. This is plausible, and although two weeks is already a reasonable time lag the effect of a longer interval between administrations might usefully be investigated. Repeated measures along these lines can, I think, make it possible to ascertain whether an informant self-coding procedure like AC tracks processing behaviours in a fairly reliable way across tasks and sessions. As the task-taker- cum-informant is herself the coder, it is to her that intra-coder reliability assessment needs to be applied.

In summary, I have tried to outline here some of the stages in the evolution of the AC procedure and its application. Although the procedure does not track processing behaviours with the subtlety that verbal report potentially offers, my suggestion is that it can record those events in cloze recovery most salient in the mind of the filler of blanks herself, and that it can do so from larger numbers of informants, if these are available, than can realistically be investigated via verbal report by a single, non-professional researcher. Data presented in the following section may offer some insight into the balance and types of processing events that the two verbal report formats and AC seem to track.

7.18 Detailed comparisons of the processing of some passage items under Think aloud, NCR and AC conditions

In Appendix 4 I show some data from think aloud and NCR protocols, and AC manuscripts, containing GL1 and JL1 informant's processing of deletions in the OLYMPICS passage. No NCR data was gathered from GL1 informants, but as only one JL1 NCR protocol was cited in full in chapter 6, it may be useful to include this data from JL1 informants for purposes of comparison with think aloud. For reasons of space it is impractical to discuss all 32 passage deletions in detail, so ten deletions were semi-randomly selected for inclusion here. Five were taken from within each of the two subsets of items outlined in chapter 4, i.e. 'grammatical', and 'lexical'. Both types of deletion are necessarily of interest in looking at how natural or fixed-ratio deletion cloze blanks come to be filled, and in this way over-emphasis on one set is avoided. The 'grammatical' deletions targeted are: (11), (13), (18), (20), and (28), and the lexical items are (2), (7), (8), (14), and (26). For each deletion I present the relevant processing information extracted from two think aloud protocols from each L1 group (along with data from two JL1 NCR protocols) and a parallel sample of AC manuscripts from G;1 and JL1 informants. Solo-condition (SC) and paired (PC) think aloud protocols were pooled together prior to selection in order to provide a larger sample (See chapter 6 for a discussion of the comparability of these data types.)

Storage of verbal report data and AC codings representing informant processing of items in a database made it possible to semi-randomly (see below) select anew for each item considered here. In this way a broader and more representative

sample of the available data is tapped than if I instead looked at the same individuals' protocols throughout: the amount of information available about the processing of deletions would then have depended on whether 'good' or 'weaker' informants had been selected, and had it been the latter then a very skewed picture–of think aloud especially–might have resulted. The selection of excerpts could not in fact be carried out in as truly random a fashion as intended, for the database software in which the item excerpts were stored allotted an individual number to each 'saved' entry. Thus, the more entries an informant's record contained, the greater the likelihood that one of her numbers would be selected.

The think aloud data is shown in Appendix 4 with only indecipherable, task-irrelevant (e.g. interaction with the researcher), or purely pair-specific verbalizations edited out, and the original passage item is shown in brackets beside each deletion. Note one change in the specifics of representation here: the informant's verbalization in English is (due solely to a software incompatibility) shown in italics, with the remainder translated and shown in regular type. The think aloud or AC informant's chosen filler is underlined. Codes selected by AC informants and identified by me in the protocols of think aloud and NCR informants are shown in boldface, with those underlined denoting behaviours or events securely identified only via post-task interviews or questions, or in the case AC informants by reference to markups or comments on their manuscripts. My observations about the data for each deletion follow the final extracts cited.

Inferences and conclusions from the data in Appendix 4

From the data excerpted we may, I think, reasonably draw several conclusions. The first is that the information provided by think aloud is not inevitably richer or thicker than that gleaned by AC at a considerably lower outlay in researcher time. The data above includes a few instances (Stella, deletions (11) and (13) Detlef deletion (13); Stephanie deletion (14), for example) in which the think aloud data tells us nothing or next to nothing about how an (often 'low-challenge' and grammatically-cued) item was filled, while AC typically offers at least some indication of how the informant did so. (This quality may be seen either as a strength of AC or as a weakness of think aloud.) NCR reporting is in general more condensed and more easily understood, but although it often performs better on easier/grammatically-cued items than think aloud, it is not richer in insight in every case. The AC data for deletion (11) is arguably as insightful as that from the two verbal report formats, and for deletion (28) it is no worse, at lower effort. It could be argued that where two formats generate much the same amount of data that requiring the lowest outlay is to be preferred. Even if verbal reporting is not inevitably more productive of insight, the table below suggests that is very often so. 'X' denotes the reporting format which I feel has produced the most insight into informant's processing in each of the ten items detailed above.

Del.	TA	NCR	AC
11		x	x
13	x		
18	x	x	
20	x	x	
28			x
2	x	x	
7	x	x	
8	x	x?	
14	x		
26	x		
Totals	**8**	**5-6**	**2**

Figure 7.14: Productivity of three reporting formats on 10 cloze items

The evaluations above reflect the overall depth or breadth of insight I see in the products of each format, but it is hard to judge this without reference to a specific goal. If we are interested in knowing only the most important cue used by informants in filling each blank, for example, then AC would be equal to verbal report in many cases. If we wish to know how often, or exactly at what point, an informant notes errors in her understanding of a passage span (and revises this) then think aloud may be the only useful format.

Something that becomes clear from the extracts above is that despite its economy in terms of researcher time see chapter 8) AC is not immune to confusing events such as Risa's coding of her choice of 'names' in deletion (14) as the product of (inaccurate) phrasal knowledge, when markings on her manuscript appeared to indicate that she had also used information in the sentence following the blank. In a sense this may be an advantage of AC, for graphic markups were intended as complementary or 'backup' indices of behaviour, as well as supplements to coding. In a situation such as this, we may infer from the markup that the

informant's coding was incomplete and amend it appropriately. (This is analogous to–and I think no more questionable than–using data from post-task interviews to gain a clearer picture of think aloud processing.) Graphic markups on AC manuscripts may also help to ascertain that a given coding was appropriately and/or consistently applied, as in deletion (8) where GL1 AC informant Vince drew a 'box' around 'took place', suggesting he saw this as a unit. The same phenomenon was found on Vince's manuscript at items (16) where 'such as' was boxed, and (26); ('set aside') All were coded as 'Phr'.

One further conclusion is that test constructors might usefully employ think aloud/NCR and AC procedures to different ends. Say that we wish to find out whether test-takers utilised in a given test item those kinds of knowledge expected or predicted by test writers or experts. AC codings and markings-up might suffice to tell us that structurally appropriate fillers were selected on grammatical grounds, or that test-takers' use of translation was restricted to 'high-challenge' items. Where events impenetrable to AC arise, or where 'easy' items prove unexpectedly difficult, then NCR or think aloud, with their potentially richer pictures of processing might be brought into play.

7.19 Post-task elaboration in AC?

But why not, one might ask, incorporate post-task questions into the AC task? Although my ambition for AC was to come up with a reporting–or, rather, a recording–format that dispensed with verbal report altogether, I did conduct brief, individual post-task interviews with four JL1 AC informants (It was not possible

to put questions post-task to GL1 AC informants.) We looked together at the coding(s) selected, and at any markups and written comments on the manuscript. For each item, I then put the same basic question with minor rephrasing:

"Can you say anything more about how you found the word for this blank?"

I took notes of any additional comments made, and informants were allowed (using green ink to distinguish on-task and post-task markups) to add comments, link passage items, etc. to their manuscripts. Two factors must be borne in mind here. Firstly, so as not to convey the idea that there would be an opportunity to pass on information orally, and thus that not everything had to be recorded on the AC manuscript, these post-task sessions had not been announced in advance. This may have contributed to the apparent unease of two out of the four informants, who may have formed the impression that they were being asked to comment further because they had not carried out the task satisfactorily. Had I not noticed this unease, and been able to reassure them that such was not the case, they might have been motivated to 'improve' matters by adding events that had not actually occurred during the AC task. Secondly, informants' affective reactions aside, the lack of any real-time record of thought processes in AC means that there is nothing to confirm or disconfirm anything the informant chooses to add, post-task unless (see 7.20, below) there is some graphic evidence to support her remarks. Of course, if an apparently salient recovery event were suddenly to be mentioned (see below) without any support, one might reasonably be sceptical that it had played a role during the task.

On the whole, post-task questions to AC informants did not seem to provide much additional insight in return for the investment of time, and I would speculate that three aspects of the task may have contributed to this. First, the lack of any announced verbal report stage (i.e. a post-task interview) arguably had not indicated to informants any need to 'prepare' or 'store' detailed verbal explanations of their actions.

Second, although the cognitive load on AC informants may be lower or less sustained than that on thinkers-aloud–in that the former do not have to keep up a continuous flow of verbalization–the requirements of the AC task (selecting the appropriate code(s); marking up the manuscript; adding written comments) may involve a fairy high level of engagement; it is possible that this does not leave much cognitive space for self-directed elaborations of the processing events being otherwise dealt with.

Third, observation of AC informants at work suggested that for many the task was, as it were, broken up into steps. Informants often (and probably typically, although my observation notes do not allow me say so with confidence) carried out the task in this order. First, passage content was read, on whatever level, and graphic markings such as arrowing, circling, etc. were made. Next, the blank was filled and, with or without reference to the set of coding cards, codings were selected and entered. Finally, difficulty ratings were entered, along with any additional written remarks. The fact that many informants (again, perhaps most) paused before adding comments, apparently to look over their efforts up to that

point, suggests a degree of separation of written elaborations from the previous sub-tasks, although it may also reflect a desire not to repeat in writing information already given by other means. It was noted that, when interrupted during orientation/practice sessions to be asked what they had just marked on their manuscripts, informants in the midst of writing comments typically had to look before answering. This too may suggest a series of semi-discrete stages in AC, each of which has to be recalled before it can be discussed.

Some useful elaborations and/or confirmations did arise in these post-task interviews, however. AC informant Yumi had filled blank (18) with 'even' (SEMAC 'also') and had circled the word 'boxing' immediately following. Yumi added that she had been surprised to see a "violent" sport like boxing mentioned (cf. NCR informant Ryou's similar remark in chapter 6) and hence had chosen 'even' as the filler. From manuscript evidence alone this would not have been clear, although the circling of boxing might have provided a clue.. AC informant Emi arrow-linked, post-task, the mention of 'foot races' following deletion (17) with her choice of 'each' for deletion (20). This she had coded only as 'Gra', implying use of local grammatical constraint in recovery, whereas the link back to 'foot races', whose plural status had, Emi said, made her choose 'each' over 'the', was evidence of longer-range cue uptake codeable as 'EP.' Although I allowed her to mark the link on her manuscript in green, I was ambivalent about this addition. On the one hand, Emi had mentioned it almost as soon as I had invited further comments, but on the other hand she had to look through her set of coding

cards to locate the appropriate 'EP' coding. One of my aims in encouraging marking up of manuscripts was that this would offer a 'double check' of processing behaviour. If an informant marked a cue–target link but forgot to code it, some evidence for it would still be there. The fact that Emi had neither coded for nor marked the cue she mentioned post-task implied that her aadded comment may not have reflected authentic task behaviour.

Of the three reporting formats considered here, only think aloud is 'linear' in the sense that the data is produced in real time and (as far as we can tell) cognitive events are sequenced chronologically rather than appearing in something like their order of salience. This is not to imply that these two principles necessarily conflict, for where only one cue commonly appears to have been activated, as in deletions (11) or (13), no conflict is possible. In NCR and AC, however, the informant is better able to select what to report, and we may reasonably anticipate that she will choose to report those events or processes which made the strongest contribution to recovery.

7.20 Other AC data: graphic markup and written comment

The cognitive burden of simultaneously recovering cloze fillers and thinking aloud about one's actions was, in the opinion of the majority of solo-condition informants a considerable one, and it would have been distracting and, for some at least, superfluous to ask them to 'add value' to their task sheets by inserting graphic devices to supplement their account of their processing. The addition of a marking-up component might have been more practical in NCR verbal reporting,

but again I did not want to distract informants from the central task of telling me what they had done. Moreover, part of my motivation for trialling NCR reporting was to see how much time could be saved by the immediate retrospection format, and adding a task beyond that or reporting would only have clouded this question.

AC was intended to be a more efficient means of gathering information about informants' processing behaviours, so that I was inclined to see whether additional tasks could usefully be attached to the central one of coding. Rating items for perceived difficulty was one such addition: the manuscript shown in figure 7.1, above, carried a difficulty rating scale for each item, and it was found that these were fairly reliably (mean 74%) filled by JL1 AC informants (and 78% for their GL1 counterparts, below.) Completion rates were likely to decline as the number of additional tasks increased, however, and it seemed advisable to make add-on tasks conform as far as possible to the kinds of thing that informants might spontaneously do in an authentic interaction with an L2 passage. It is, for example, common for Japanese L1 students (at least those educated within Japan) to link with lines and arrows semantically related items within a text, examples of similar syntactic structures, etc. Key information, or spans of text which are found challenging are often underlined in pencil, and this underlining (informants have told me, and my own experience of high school teaching confirms) is often accompanied by translation into the L1.

If such devices could be co-opted to supplement coding choices, it might, I anticipated, be possible to expand the range of events open to tracking via AC.

The abovementioned risk that AC might fail to stimulate reports of difficulty with extant passage content might be overcome by gaining insight into informants' use of translation, which I hypothesised might be seen as a reflection of difficulty in comprehending passage content. The suggestion that AC discourages informants from 'co-processing' deletions might be addressed by providing them with some way to link passage items which were in some way connected during processing. It seemed realistic, then, to encourage AC informants, as part of their task, to do the following (These devices were illustrated on an OHT during orientation):

(1) to mark by underlining or by boxing or highlighting any passage content which they found themselves translating, the choice being their own, provided this was consistently applied
(2) to link via lines or arrows blanks and any spans of passage content which were relevant in filling these, including links between fillers if this were appropriate.

The habit of some informants, mentioned above, of using underlining to identify important points in the passage made it necessary to agree on alternative devices by which they might mark anything they wished to. Informants who did not feel able to reserve underlining solely for translated content were encouraged to use another of the devices above for this purpose (In the discussion and representation that follows, underlining is used to stand in for these other devices, too.) and to make me aware of which they had chosen: in this way I could identify the spans of passage content the informant had marked as translated. I discuss identified graphic links below, and then take up the contribution of written comments. The evidence for translation provided by AC manuscripts is given a separate section,

below.

Graphic links between passage items

Links identified on at least 10% of informants' (i.e. three manuscripts of the (32-deletion) OLYMPICS were:

(a) national <--> international (deletion (3))
(b) no (deletion 11) <--> a (deletion (12))
(c) blank (21) <-->> (boys') gymnastics / discus and javelin (throwing) / foot races
(d) blank (21) <--> boxing and wrestling
(e) fourth day (preceding deletion (25)) <--> blank (27)

Given that it took some time to spot and formulate questions about graphic markups, like any other processing event, it was not always possible to take these up with AC informants post-task (but see below.) 'Educated guesses' thus had to be made. Of the links noted above, (a) very likely reflected the collocational link commonly identified between these items in think aloud data, while (b) almost certainly indicated informants' recognition of these two blanks as belonging to a learned grammatical structure (cf. the protocols of Yasuko and Ryou) and as such may have reflected much the same linked processing as was noted in think aloud, while (b) and (c) arguably signalled uptake of cues from an earlier sentence or paragraph. As only one of the three informants whose manuscripts bore link (c) had in fact coded deletion (21) with 'EP', this particular link may indicate a secondary cue to the 'same sentence cue' codings ('SSf'; 'SSb') entered by all three. It is plausible that the local context told these informants that deletion (21) had to be filled by a sport, while reference back to those mentioned earlier told

them which it was likely (not) to be.

Idiosyncratic graphic links were too numerous to describe here, but one such link identified on two JL1 manuscripts was taken up post-task with one of the informants, and this example may illustrate how graphic links can help recall the chain of reasoning involved in a recovery and/or coding—even though in this case the filler is not semantically appropriate. For OLYMPICS item (20), the two manuscripts carried the filler 'each', and bore the codings 'Gra' and 'SSb' (cue earlier in the same sentence.) Both also bore a graphic link between items (20) and 'special tests'. The informant I questioned post-task indicated that the plural status of this phrase had prompted her to fill item (20) as she had. As on many other manuscripts (see 7.21, below) 'pentathlon' had been underlined, and the informant confirmed that she had not understood the event's 'unitary' status. Apprised of this, she realised at once that the only appropriate filler for (20) would be 'the,' A mistaken cue uptake, then, appears to have triggered the incorrect filler, but a lack of lexical knowledge may have contributed to the error.

Supplementary written comments in AC

The instruction to AC informants to use the space available associated with each deletion to add any supplementary comments they wished to make was less productive than I had anticipated, for my hope had been that the absence of a sustained reporting requirement would free AC informants to elaborate in some detail about their processing. Putting aside for now the question of how useful the comments were, and going by raw counts of items for which a comment was

entered, we find that JL1 AC informants entered a comment in 198 out of 768 possible instances (n=24 x 32) or just over 26%, with 58% pertaining to lexical items. GL1 informants entered comments in 185 out of 512 possible cases, or just over 36%, and of these just over 67% pertained to lexical item deletions. By these raw counts alone, it would appear that GL1 informants made considerably better use of the provision for written elaboration, but the picture changes somewhat when we look in more detail at what was entered by the two groups.

Some 80% of JL1 comments, and just over 85% of GL1 entries, were mainly or solely in the L1. This may imply something about the informants' use of the TL and L1 as their main language of processing, or it may simply reflect the fact that it can only be eaiset to write comments in one's own first language.

	JL1 (n=24)	% relating to lexical; structural items	GL1 (n=16)	% relating to lexical, structural items
% of deletions commented on	26%	58%; 42%	35%	67%; 33%
Overall % of comments in L1	80%		85%	

Figure 7.15: Comments entered on AC manuscripts

Comments by both informant groups could be (with some comments, especially by JL1 informants (see below) separable into two or more components) roughly allocated to the headings below, although a small percentage fell into grey areas and had to be left out of the classification. 'Affective' comments related to perceptions of item difficulty (which had typically also been rated on the scale

provided) and/or statements of (usually low) confidence in the chosen filler ("*Bestimmt falsch!*" / "Surely wrong!", etc.); 'repetitive' entries essentially repeated information already given by the informant's selected coding(s), as in coding 'Gra' accompanied by "*Bunpo mondai*"/ "[It's a] point of grammar.") 'L1Eq' entries appear to have represented attempts to demonstrate some understanding of the missing items' meaning where a SEMAC correct filler might or might not have been entered in the blank. These entries were perhaps more interesting when they revealed a misunderstanding of the passage content on the informant's part, but as no 'new' misinterpretations were identified on AC manuscripts these are not discussed here. 'Informative' comments (so labeled because they conveyed information not aleady provided by other elements of the AC task) contained a wider range of meaning. Some indicated uncertainty over which coding was more appropriate ("Phr? Col?") or offered alternative filler items ("Can 'rulers' go here?")

	'Affective'	'Repetitive'	'L1 Eq.'	'Informative'
JL1	40%	21%	34%	26%
GL1	22%	24%	46%	16%

7.16: A rough breakdown of informants' comments by 'function'.

As some entries could be broken down and the elements placed under more than one heading, percentage totals may add up to more than 100. A chi-square analysis of the raw frequencies with which entries were allocated to the headings above produced a value of 9.9345, significant at 0.05, df3. The written comments

of Jl1 and GL1 informants may thus be said to differ significantly in terms of the functuons attributed to them here.

One further 'functional' type of written entry on AC manuscripts remains to be discussed, and was produced overwhelmingly (mean 2.1 per informant) by JL1 informants. These *kanji* were often written outwith the space provided for elaborative comments, and did not appear to be aimed at the researcher. L1 glosses of TL words or key phrases (e.g. '月桂感' *laurel crown*) were often added interlinearly, and the passage content they glossed or summarised was in most cases also underlined (though these L1 renderings were evidence of translation, even in the absence of underlining.) Some longer L1 entries did appear in the comment 'boxes' but these too appeared to be memos to the informant herself, as in: 女は出席出来ない' *women cannot [sic] attend.* One AC informant noted, post-task, that she had not at first realised that the pentathlon was a single, multi-stage, event; hence her reminder to herself of '色色の競技' (*various games*.) The few single word or phrasal glosses which were entered in the boxes were taken to represent 'L1Eq' entries, as for GL1 AC informants.

To sum up this discussion of comments in AC, perhaps the two key points are that informants produced rather fewer comments than I had hoped to gather, and the bulk of these offered little or no new insight into informant's processing. With regard to the first point, post-task questions revealed that while JL1 AC informants may not have been sure in their own minds about the function of the comments space, a higher number indicated that they had not added comments

because they had nothing important to tell. As for the second point, the percentage figures above offer few surprises. It is fairly well-known to teachers and testers in Japan that students in that country often present themselves as having little confidence in their performance. The structural similarity of the TL and German may have enabled speakers of the latter language to offer more direct single-word glosses for TL items, while the comparative economy of Japanese *kanji* script as a medium of transmission may have allowed its users to construct 'denser' comments than their GL1 counterparts, with a similar outlay of time and effort. The scarcity (mean 0.6 per informant) of apparently self-directed memos on GL1 AC manuscripts may reflect that culture's lower proclivity (in my own teaching experience) for writing on textbooks and printouts. On a more positive note, the provision of space for comments appears to have allowed informants a means of expressing affective reactions to the task, and of course a means of demonstrating understanding of item meaning even where a TL filler was not achieved. This 'face-saving' aspect may be, as I argue in the concluding chapter, of considerable importance even in introspective tasks as superficially restricted as AC.

7.21 Translation in AC & predicted item difficulty

As may be clear from the discussion of protocol data and post-task retrospection in chapter 6, neither think aloud nor NCR reporting formats seemed to provide reliable indices of how extensively the L1 was used in cloze processing. The absence of L1 as a language of processing (as opposed to one of reporting) in NCR was to be expected insofar as this format requires the informant to report

post facto her fillers and the steps by which she arrived at them, but not the medium in which this was done. The incomplete picture of L1 use provided by think aloud is due not just to the built-in assumption (Ericsson & Simon 1984/1993) that even think aloud verbal report cannot reveal more than a fraction of what an informant was doing at any given point, but also because some processing in the L1 appears to be consciously or unconsciously 'suppressed' in that procedure.

It is not altogether clear why some. But not all, think aloud informants often (but not always consistently) seemed to speak lower and less distinctly when using their L1 (In a number of instances of '[INAUDIBLE]' speech, native-speakers of the informant's L1 whom I consulted were able to ascertain that she was using her L1, although not what she was saying. As mentioned in chapter 5, it was not realistic to query every instance of inaudible protocol data in a post-task interview.) but my speculation was that this shift reflected greater perceived difficulty or uncertainty in processing passage content and/or recovering fillers. In such a situation, it would be reasonable to expect that informants' verbalization would become more hesitant and more self-directed, such that oral production might temporarily decline. Processing in the TL at such times might largely take the form of (perhaps repeated) readings of the span of passage in focus, and support for this notion can be found in the think aloud protocols of both GL1 and JL1 informants. GL1 think aloud informant Detlef confirmed post-task that the extended inaudible span of verbalization below had been in his L1, and had

represented an attempt to construct an equivalent meaning in German for the passage context of OLYMPICS item (37):

"[...] worthwhile.. the training [INAUDIBLE ca.12 secs] the public honour also made the first discipline of the ten month training worthwhile.. I think this is quite hard now.. public honour okay public honour made.. I'm not so sure about the first but hmm.. public honour also [...]"

We have seen from think aloud data in chapter 6 that informants may mention L1 equivalents or parallels in the forms of lexical items or syntactic structures. Some of these events appear to represent the informant's wish to demonstrate her understanding of the context of the blank even though she is unable to supply a TL filler. L1 glossing of isolated lexical items such as 'waived'; 'dishonoured', 'javelin', etc., may merely represent a kind of 'translation-in-passing', incidental to recovery of any blank, but it is unsafe to rule out any recovery role altogether for such behaviours. Post-task, informants may or may not be able to recall their reasons for glossing an item in their L1; for some it may well have been a reflex response to encountering an unfamiliar item, but it does not seem likely that informants were able, in 'real time', to know what would and what would not be relevant to recovery of upcoming blanks. It is likely that the very 'embeddedness' of cloze items often makes it hard to be sure whether translation is being applied directly to the recovery of an individual blank, or whether it is aimed at understanding extant passage content. It may be unrealistic to even attempt a dichotomy like this, for in many cases, informants interviewed post-task–or even interruptively and within moments of translation being used-were unable to say

what the goal of their translation behaviour had actually been. Given that translation of more than a few consecutive words of a natural cloze passage cannot help but at least 'co-target' a deletion, the difficulty of distinguishing functions of translation (cf. Krings 1988) will be clear.

My hypothesis was that recourse to translation (I use the term here to cover verbatim translation and L1 paraphrasing) of a longer span might be taken as an indication of difficulty with its content and/or the filler(s) it contains. I suggest that AC, with its instruction to informants to underline (or mark via another convention) any passage content translated, may be a less ambiguous index of what gets put into the L1 than think aloud, and hence of which spans of text and/or items informants found challenging. We might reasonably expect to find a good deal of 'wild', or unsystematic, translation on AC manuscripts, according to how familiar individual informants were with passage lexis, grammatical constructions and even aspects of the topic. If AC is a useful index of translation-cum-difficulty, however, we might also expect to find more consistent underlining of those passage contexts which native-speaker consultants predicted would be difficult for informants from their own language backgrounds. If (even though JL1 consultants rated item difficulty higher overall) we may take a value of '3' or above on the five value scale as indicating for both L1 groups the perception of above-average difficulty, we see the following OLYMPICS passage items ranked as challenging: by GL1 consultants

(7); (9); (15); (19); (21); (23); (25); (26) and (29)

and by JL1consultants

(1); (2); (7); (9); (15); (19); (21); (23); (25); (26) and (29)

The 82% overlap between the two sets may or may not be interesting (I would have anticipated that JL1 consultants would have outpaced their GL1 counterparts by more than this in marking items as 'difficult'.) but anyway suggests similar loci of difficulty in cloze for individuals from language backgrounds as distant as German and Japanese. Apart from the overall higher modal values attributed by JL1 consultants, differences exist at the individual item level. GL1 consultants ranked item (1) at the low end of the scale, while JL1 consultants ranked it as '3'. Both groups recognised that successful filling here depended on extratextual knowledge, and we may assume that GL1 consultants simply expected this to be more widely held than did their JL1 peers. The gap between GL1 and JL1 ratings of item (2), '2' and '5', respectively, may reflect the arguably more direct equivalence of English and German constructions over that of the TL and Japanese.

How best to 'depict' translation behaviour on the page is not obvious, for even where individual phrases such as those above are underlined in isolation, as it were (perhaps with SEMAC fillers before and after) it is hard to know how much of a role that translation played in recovery of the adjacent deletions. Rather than attempt to distinguish translations more and less salient to recovery, I have opted to try to depict translation behaviour as follows. It would be next to impossible to distinguish on the page the 16 (for GL1AC informants) potential frequencies of

translation for a given span, let alone the 24 for JL1 informants. Instead, four levels are shown. Passage content underlined by fewer than ca. 25% of the informant group (i.e. GL1 one to four informants; JL1one to six) is shown in regular type. That underlined by 25%-50% (GL1 five to eight; JL1 seven to 12) is in boldface. Content underlined by 50=75% (GL1 nine to 12; JL1 13-18) is in bold capitals, and that underlined by75% and above (GL1 13–16; JL1 19-24) is in bold, underlined capitals. (Fillers are not shown here, but where present on informants' AC manuscripts these were typically included in the spans underlined.) The resulting versions, below, show that (on a crude count of words underlined) German L1 AC informants underlined less, overall, than their JL1 counterparts (88 vs.119 words underlined by more than half of informants, 42% and 57% of the passage extant word count of 208), respectively) and while this lower figure very may indicate a lower overall need to translate passage content, it may also reflect a lower baseline of educational/cultural predisposition to do so: underlining in textbooks seems to be considerably less common in Germany than in Japan, where pupils need not return books issued, and marking and glossing of unfamiliar or pedagogically salient items is the norm

UNDERLINING BY GL1 AC INFORMANTS (N=16)

In ancient Greece athletic festivals were very important and had strong religious associations. The Olympian athletic festival, held every (1)_____ years **in honour of Zeus, EVENTUALLY** (2)_____ **its local character,** became first a (3)_____ event, and then, after the rules (4)_____ foreign competitors **had been WAIVED**, international. (5)_____ **one knows exactly how far back** (6)_____ **Olympic Games go**, but some **OFFICIAL** (7)_____ **DATE** from 776 B.C. The Games (8)_____ **place** in

August **on the plain** (9)_____ **Mount Olympus.** Many thousands of spectators (10)_____ from all parts of Greece, **but** (11)_____ **married woman was ADMITTED even as** (12)_____ spectator. Slaves, women and **DISHONOURED PERSONS** (13)_____ not allowed to compete. **THE EXACT** (14)_____ **OF EVENTS** is uncertain, but events (15)_____ boys' gymnastics, horse-racing, **field events** (16)_____ as **DISCUS** and **JAVELIN THROWING**, and (17) _____ very important foot races.

There was (18)_____ boxing and wrestling and **SPECIAL TESTS** (19)_____ **VARIED ABILITY** such as the **PENTATHLON**, (20)_____ **winner of which EXCELLED in** running, (21)_____, **discus and javelin throwing** and wrestling. (22)_____ evening of the third day was (23)_____ **TO SACRIFICIAL OFFERINGS to the heroes** (24)_____ **the day**, and the fourth day, (25)_____ **OF THE FULL MOON, was SET** (26)_____ as a **holy day**. On the (27)_____ and last day, all the victors (28)_____ crowned with **HOLY GARLANDS OF WILD** (29)_____ **from a SACRED WOOD**. **So great** (30)_____ **the honour** that the winner of (31)_____ foot race **gave his name to** (32)_____ **year of his victory!**

Underlining by JL1 AC informants (n=24)

In ancient Greece **athletic festivals** were very important and had **strong religious associations**. The Olympian athletic festival, held every (1)_____ years **in honour of Zeus, EVENTUALLY** (2)_____ **ITS LOCAL CHARACTER, became first** a (3)_____ event, and then, **after the rules** (4)_____ **foreign competitors had BEEN WAIVED, international.** (5)_____ one knows exactly how **far back** (6)_____ Olympic Games go, but **SOME OFFICIAL** (7)_____ **DATE FROM** 776 B.C.

The Games (8)_____ place in August **on the plain** (9)_____ Mount Olympus. Many thousands of **spectators** (10)_____ from all parts of Greece, **but** (11)_____ **married woman WAS ADMITTED even as** (12)_____ **spectator. Slaves,** women and **DISHONOURED persons** (13)_____ not allowed to compete. **THE EXACT** (14)_____ **OF EVENTS is uncertain but EVENTS** (15)_____ **boys' gymnastics,**

horse-racing, field events (16)_____ as **DISCUS and JAVELIN throwing**, and (17) _____ **very important FOOT RACES.**

There was (18)_____ **boxing and wrestling and SPECIAL TESTS** (19)_____ **VARIED ABILITY** such as the **pentathlon**, (20)_____ **winner of WHICH EXCELLED IN** running, (21)_____ , **discus** and javelin throwing and wrestling. (22)_____ **evening of the third day was** (23)_____ **TO SACRIFICIAL OFFERINGS** to the heroes (24)_____ the day, and the fourth day, (25)_____ **OF THE FULL MOON**, was **SET** (26)_____ **AS a holy day**. On the (27)_____ **and last day**, all the **victors** (28)_____ crowned with **holy GARLANDS OF WILD** (29)_____ **FROM A SACRED WOOD. So great** (30)_____ **the honour** that the winner of (31)_____ foot race **GAVE HIS NAME to** (32)_____ **YEAR OF HIS VICTORY!**

7.18: Frequencies of underlining of passage content by Gl1 and JL1 AC informants.

I will choose, arbitrarily, a frequency of underlining of 50%+ as indicating hypothetical serious difficulty among the informant pool as a whole. By this criterion (and that of 3+ rating on the consultants' scale, above) GL1 and JL1 informants translated passage content related to the items below:

GL1predicted (above) & observed (below) difficult items

		7	9		15	19	21	23	25	26	29	
	2	*				*		*	*	*	*	

JL1predicted (above) & observed (below) difficult items

1	2	7	9		15	19	21	23	25	26	29	
	*	*		14	*	*		*	*	*	*	32

Figure 7.19: OLYMPICS cloze items predicted by GL1 & JL1 consultants to be difficult , and actual difficulty as reflected in translation of surrounding context (* indicates match between predicted item difficulty and observed underlining of associated passage spans.)

German L1 consultants, then, accurately predicted difficulty as reflected in underlining /translation for only six of nine items, or 66%, and failed to predict difficulty with item (2). Their Japanese L1 counterparts predicted 8 of 11, or 73%, but failed to predict difficulty with items (14) and (32). I would speculate that the slightly greater predictability of item difficulty by Japanese consultants (like their GL1 peers, mostly educators of one kind or another) was occasioned by Japan's more centralised and uniform EFL curriculum, and the higher proportion of their total English exposure that Japanese students obtain from that curriculum. But is there a further association between higher difficulty ratings ('4' or '5' on the scale) and 75% underlining of passage content? Items rated at '4' or above on the scale by GL1consultants were (23); (25); (26) and (29). Underlining at the 75% level is, however, only founding relation to item (29). JL1 consultants rated items (2); (7); (15); (19); (21); (23); (26); (29) at '4' or '5' on the scale, but 75% underlining was found only for the contexts of items (2); (19); (25) and (29). By this criterion, both consultant group accurately predicted higher difficulty for 25% of the items so rated. There appear to be limits, then, to what L1 consultants' intuitions can tell us about a task like cloze: predicting which items on a task will be difficult seems to be easier than predicting how difficult these will be.

I hypothesised above that translation behaviour could be seen as an indication of difficulty extant passage content and/or recovery of related fillers. This notion requires some modification in the face of the lower (i.e. below 50%) but still unexpectedly high frequencies of translation of predicted low difficulty items such

as (30) for both L1 groups and (5) for GL1 informants. GL1 think aloud informant Claudia recognised the necessary filler for (30) at once, adding later that she had learned it in school. Some GL1 informants read aloud this passage span in their L1, apparently without effort. Here Tom reads first in English and appears to suddenly note the exact L1 parallel form:

"[...] so great was the honour.. ach so [...] gross war die Ehre [..]"

On one level this does represent a translation into the L1, but it follows the successful filling of the blank and hence cannot have played more than a confirmatory role in recovery. It is possible that some GL1 AC informants underlined this span as translated more because it are so readily 'translatable' via the parallel construction in the L1 than out of any difficulty in filling the blank. The case of JL1 translation is perhaps more problematic, for to the best of my knowledge, Japanese lacks any such close parallel construction. The TL *touch*i ('inverted') construction is, however, taught in school English lessons, and is typically quite familiar to Japanese learners. Despite this, JL1 informant Yasuko appeared to take a few moments before getting the filler on her third 'pass'

"[...} so great SOMETHING the honour.. so great the honour.. so great was the honour [...]"

but did not overtly translate. Why this span should have been underlined by more than 25% of Jl1 informants, however, is unclear, for although each of the four JL1 consultants to whom I showed the blanked sentence *"So ..... was the honour that the old poet cried tears of joy"* was able to fill the blank almost at once, none could immediately come up with a Japanese equivalent. If this apparently familiar construction with its readily fillable blank (78% acceptably recovered by JL1

informants) was in fact translated by those who underlined it, it would not appear to have been a task-wise use of their time.

In summary, then, although recourse to L1 translation seems to be applied to 'difficult' spans of passage content and to deletions they contain, easier content may also be glossed on some level because doing so does not take up too much processing effort, as in the TL-German parallel construction discussed above. It appears to be a staple view among translation specialists (Krings 1988) that translation into the L1 is multifunctional. One the one hand, it may aid comprehension of otherwise uninterpretable content, and on the other it may make it easier to access meaning in a more condensed form and/or at a more global level.

If these various levels of functions apply in reading of conventional L2 texts, it is quite plausible that they may also be applied to cloze, even though the evidence from AC manuscripts suggests that L1 glossing is driven more by the local context surrounding more challenging cloze items, and perhaps also to some extent by a kind of 'recognition' factor associating passage content with comparable L1 structures and/or learned TL material. If, as think aloud data from cloze task-takers sometimes suggests, translation plays a rather larger and/or more complex role in natural cloze processing than the underlying cloze construct of linguistic redundancy might predict, it would be useful to find out more about how and when it comes into play. On the other hand, I have not identified any clear association of greater or lesser recourse to translation with cloze success, and

to that extent the question may not appear urgent to the real-world application of cloze as a measure.

7.22 Reporting format and cloze success

The relative difficulty of individual cloze items can be ascertained simply by setting the task to a sufficiently large pool of test-takers. While verbal report can give us some insight into how much work a test-taker has to do to fill a blank, and what kind of work she does, the small sample sizes and possibly confounding variable of reporting format (think aloud vs. NCR) do not sit well with objective measurement of item facility. Fortunately (although I have argued in chapter 5 that at least some of my informants appeared to treat the verbal report task much as they would have an authentic test) the real-world overlap between 'test-like' data-elicitation tasks and actual tests is precisely nil. I was thus less interested in how well my informants scored on the stimulus cloze passage they completed than on how they arrived at fillers, whether SEMAC or not. That said, the cognitive load imposed by reporting may impact on processing of the stimulus task: cognitive resources taken up by *x* cannot after all be available for purpose *y*. (I have suggested in chapters 5 and 6 that this cognitive load is greatest in think aloud procedure, although to some extent the lack of any time limit on the think aloud task may negate any effect of higher cognitive burden.)

Although our main concern is with whether and how different formats affect the reporting of processing behaviours, might there also be some value in knowing a data-elicitation task impacts on success on the stimulus task itself? This is not

simply an abstract concern, as I will explain. It appears to be a standard ethical guideline in research (S. Hammond, pers. comm.) that participants be allowed to withdraw *at any point in the research project.* To date I have had a number of apparently capable think aloud informants ask to withdraw during the post-task interview, or even immediately following their data-elicitation session. Of these, four JL1 informants could not be dissuaded from withdrawing, resulting in the loss of their audio-protocol data. These informants' motives for withdrawing at such a late stage were not altogether clear, but in at least two cases it appeared that the informant's perceived or actual lack of SEMAC success influenced her decision. In a 'snowball' sampling procedure, moreover, informants are less likely to recruit others for a task which they themselves found frustrating or discouraging. Although in practical data-gathering situations a format's impact on score may have to be weighed against its advantages or shortcomings in terms of productivity, it is at least worth asking whether reporting conditions (here, pair- vs. solo-condition, and think aloud vs. NCR vs. AC) can be shown to have an impact on informants' cloze scores–scores being something in which my JL1 informants have repeatedly shown considerable interest. Below I present some findings from t-tests applied to cloze scores, along with means in various conditions as shown.

	SCTA means	PCTA means	NCR means	AC means
GL1	74% (n = 8)	71% (n =16)	n/a	75% (n =16)
JL1	67% (n = 8)	69% (n =16)	73% (n =10)	68% (n = 24)

Figure 7.20: Informants' mean scores in 4 formats

Think aloud in solo- and pair-conditions

T-tests (SSP v.2.0) revealed no significant difference between the scores of the full sets of GL1 pair-condition and solo-condition think aloud protocols (t=1.9231; ns. at 0.05, df22) but significant difference between the solo- and pair-condition scores of their JL1 think aloud counterparts (t=5.2923 sig. at 0.05, df22.) There is evidence, then, that the paired or solo-reporting format affected the success of one L1 group but not the other.

(Solo-) think aloud vs. AC

A t-test applied to solo-condition think aloud and AC scores revealed no significant difference between means for GL1 informants (1.8653; ns. at 0.05, d22) Nor was a significant difference revealed for JL1 solo- think aloud vs. AC (t=1.7538; ns. at 0.05, df46). There is thus no reason to think that the choice of think aloud or AC task conditions impacted on cloze success for either L1 sample.

Think aloud vs. NCR

A t-test applied to think aloud vs. NCR scores did produce a significant difference in means (t=2.5240; sig. at 0.05 df32) but, as I have already noted in chapter 6, the NCR informant group contained a significantly larger number of members with experience of living outside Japan, and was thus perhaps not directly comparable in terms of TL proficiency. No wider conclusion can safely be drawn about the possible effect of NCR format reporting until a larger sample can be gathered via that format.

Reporting formats vs. naturalistic cloze conditions

The figures above are useful in that they suggest a rather limited impact of reporting conditions on task success as measured by scores, but it also seemed useful to compare the results above with those obtained from cloze taken under naturalistic conditions (i.e. without any requirement to report) in order to see whether any format was closer in terms of scoring to the (semi-) authentic cloze task. This, of course, required that the comparator cloze passage be presented either as an authentic test, or as a 'test-like' didactic task. No opportunity presented itself to administer an authentic cloze test to the available pool of students, so that classroom/didactic presentation was the only option. Owing to the nature of the classes I had been scheduled to teach, however, no intact group closely comparable (by the criteria detailed in chapter 5) to the earlier JL1 informant groups was available to complete the same cloze passage under naturalistic conditions.

Two intact groups were set the OLYMPICS cloze passage under similar and fairly generous time constraints (i.e. limited only by class time available) but for both groups mean scores were almost 20% lower than those for the informant groups above. It is counter-intuitive that the absence of any reporting burden should lead to reduced scores, and these must have been due in part to the wider range of TL proficiencies represented in the two intact groups. Perhaps the main factor underlying the lower scores of the intact class groups, however, was the high percentage (mean 22%) of blanks these groups' members left unfilled. I would

speculate that, not having volunteered to assist in data-gathering, and having (unavoidably) been told that this didactic activity would not count towards semester grades, these students were less than highly motivated to complete the task. Direct comparison of (ideally the same) student-cum-informants' success on naturalistic cloze and cloze+verbal report using comparable tasks remains an area for future research.

7.23 Effectiveness of AC as a data-elicitation tool

In this section I discuss the productivity of AC as a means of eliciting data about cloze passage processing, and compare this to those of the formats discussed in chapter 6 (think aloud and NCR) and in chapter 7, above (questionnaire items.) The measures of productivity compared are: (1) comprehensiveness of coverage, in other words, the percentage of blanks for which each format provides an indication of at least the most salient processing event contributing to recovery, and (2) depth of coverage, by which I mean the extent to which a format provides insight into events than the most salient. This will involve comparison with the two verbal report formats discussed in earlier chapters, although, as outlined in chapter 6, some processing events are taken to be more amenable to tracking in one format than in another; comparison across reporting formats then only makes sense for events which appear to be fairly equally recordable by all.

(1) Comprehensiveness of coverage

What I look at here is simply the percentage of informants in each reporting condition who provided enough information for each cloze item to allow me to apply in the two verbal report formats one or more codings to describe their

recovery or attempted recovery or who offered one or more codings of their own in AC. The table containing these percentages for the full 32-deletion task passage necessarily had to be split over three pages, and is shown at the end of this chapter; the essential details are summarised below.

Given that the data provided by informants in the two verbal report conditions might allow for no coding of their processing behaviours, one coding, or more than one coding, the easiest figure to go by is that denoted by '???', which indicates the percentage of informant protocols for which no coding was possible for an individual item. The same is true of AC, except that here the figure indicates the percentage of informants who omitted to provide their own coding for the item. The overall mean percentages of uncodeable / uncoded items for the three reporting conditions are:

Think aloud	NCR	AC
15.3%	3.8%	5.1%

Figure 7.22 : uncodeable/uncoded items in TA, NCR & AC (%)

The figures above suggest that NCR is considerably more reliable than think aloud in providing sufficient data to allow coding of processing events, with AC a fairly close second. A one-way ANOVA analysis (SSP v.2.0) conducted on the percentages of uncodeable items in think aloud, NCR and AC provided an F value of 12.60, indicating significant differences among the three reporting conditions. If we separate out structural/functional items (above) and lexical/content items (below) as discussed in chapter 5 we get the following percentages:

Think aloud	NCR	AC
20.2%	4.2%	6.0%
11.1%	2.6%	4.6%

Figure 7.23: uncodeable/uncoded items in TA, NCR & AC

(structural above; lexical below)

T-tests (SSP v.2.0) applied to means of uncoded / uncodeable items revealed a significant difference between the structural and lexical means for the think aloud (t= 2.7432; df30) condition but not for NCR (t= 0.6886) or AC (t= 0.7405.) At least in percentage terms, however, we see that structural' items, which are claimed (Oller & Jonz 1994) to be generally easier to recover than lexical items, are less frequently coded in all formats.

Why, then, does think aloud seem to less reliably provide data that allows coding of informants' processing, and in particular their processing of structural items? As processing time for think aloud informants was essentially unlimited, thus to some degree countering any 'slowing down' of processing due to the burden of continuous verbalization, the difficulty posed by recovering appropriate filler words for these items should not have been a factor: this should have been approximately the same across all three conditions. I would speculate that part of the explanation lies in the comparatively uninterrupted flow of think aloud, which fixes no 'discrete' points at which is carried out, as is the case in NCR or AC. This aspect of think aloud is at once a strength, allowing for the linking or co-processing of items (cf. the protocols of Claudia and Yasuko in chapter 6) and

a weakness in that it is arguably easier for think aloud informants to lose sight of the reporting dimension of the task. The more frequent omission of reports about structural items may be explicable if we assume that 'easier' items (i.e. items which require less or less conscious processing effort) are more likely to be recovered quickly and thus to be passed over for reporting. By making a clear distinction between the 'thinking/recovery' and reporting aspects of the task, on the other hand, NCR and AC may help the informant to balance the two, and plausibly make it more difficult to overlook reporting of even those fillers which were recovered quickly and/or easily.

While a higher probability that some processing events will go unreported is a mark against a format, it may be that loss of information about the processing of structural items is less serious than omission of information about recovery of lexical items. What seems more likely to get lost in think aloud is the uptake 'in passing' of immediate-context cues based on grammatical knowledge and, to a lesser extent, knowledge of units (e,g, 'such as') whose status as phrases or grammatical constructions (Some have coded this item as 'Gra' and others as 'Phr'.) may be unclear to informants. Articles are one class of item more likely to be recovered without any explanation, as are auxiliary verbs.

It could be argued that we do not really need to be told how an informant recovered the filler in item (13), say ("..dishonoured persons <u>were</u> not allowed to compete..") as this could only have been done through knowledge of TL structure. This recovery does not involve more, or less, salient cues, but rather only one

possible cue; the fact that the informant chose an appropriate filler means that she *must* have used knowledge of TL grammar, whether or not she made this explicit. This is plausible up to a point, but the question then arises of how widely this exemption should apply, The protocol extract below contains no explicit information about the kind of knowledge the informant used to fill OLYMPICS item (5), yet she did so correctly:

"[...] international... NEXT.. no one knows exactly.. PROBABLY [...]"

If use of grammatical knowledge can be assumed, even without explicit reference, to be the only possible route to recovery in item (13), above, can we safely assume that (equally 'local') phrasal or collocational knowledge drove the recovery in (5)?

To sum up on this discussion of 'missing data', the lack is more problematic in some cases than in others. Accepting that what may well be the most salient event in the recovery of a filler can go unrecorded in a data-elicitation task arguably represent a first step on a slippery slope, however, so that the comprehensiveness with which a reporting format records processing events may be no small issue. On the face of it, the figures above suggest that the most reliable choice in terms of range of coverage would be NCR, but this still has the drawbacks of requiring recording facilities (although, as I suggested in chapter 6, the availability of cheap and portable digital recorders to some extent liberates the gathering of oral data from the traditional venues of researcher's office and language lab) and–unless it is deemed acceptable to code directly and solely from the audio-recording–for

transcription. The preparation of NCR data requires considerably less time, however. To take one example, the transcription and coding of think aloud informant Yasuko's think aloud data required (matched for coverage of 32 items) approximately 40% more time than that for informant Ryou's NCR protocol data (chapter 6) even though both made extensive use of the TL in reporting and (in my view) reported in a comparatively clear and structured fashion within the constraints of their respective formats. The greater structuredness or 'other-directedness' of NCR data was central to this economy of time and effort.

(2) Depth of coverage, or what does AC 'lose'?

I suggested in chapter 6 that the choice between concurrent think aloud procedure and the immediate retrospections of NCR would depend in large part on what kinds of data one hoped to recover, as well as on how much time would be available for processing and analysis. Examination of the think aloud data reported in chapter 6 and appended to this paper will show that 'richer' spans of protocol data can be mined extensively for identifiable events, although it is by no means clear how salient many of these actually were in the central task in cloze, the recovery of a suitable filler word for each blank. Nevo 1989 intuited that some behaviours (such as elimination of mismatching distractors) contributed to successful answering of multiple-choice items while others were thought not to contribute. 'Guessing blind' was one such, but while favouring 'guessing' over seeking cues in the passage would clearly be a bad idea in any text-based task, guessing as a last resort may be another matter. On the level of the individual

informant-item interaction, I would argue, the role of a given processing operation or event may be nuanced and difficult to interpret for the informant herself, let alone a third party.

Nevo (op.cit.) also claimed that her checklist approach allowed her informants to identify the 'primary' and 'secondary' strategies they employed in choosing the best answer in a multiple-choice test of reading. Can AC do the same for cloze processing? It bears repeating here is that it is not always clear which processing events contributed to recovery, and by how much, and which were–for want of a better term–incidental. Even if it is impractical try to calibrate the relative contribution of every event instantiated in the data, however, we may form some basic hypotheses. Firstly, where this distinction can be made with confidence, we may reasonably assume that a processing event that preceded or accompanied a recovery was likely to have been more salient than one that followed:

"[...] OKAY.. I WOULD SAY HERE no one knows how far back.. THAT'S A WELL-KNOWN EXPRESSION [...] AND HERE IT SAYS but SO [the preceding part] MUST BE NEGATIVE.. no one [...]"

It is also reasonable to suppose that a clearly more frequent behaviour or event is likely to be more central to recovery than a less common one. In OLYMPICS item (20) above, grammatical knowledge quite clearly played the greater role, while cues in the earlier paragraph contributed to recovery of an inappropriate filler. In item (12) the key cue is again 'Gra', while only a very few informants also possessed the extratextual knowledge that may have confirmed, or to some extent helped in, the recovery. This problem of deciding which events contributed

most to recovery only presents itself in think aloud, for in NCR and AC formats the informant herself makes the decision for us. As exemplified in the protocol data shown in chapter 6 and Appendix 4, NCR informants appear to 'condense' their reports to what seems most salient, while in AC the norm appears to be to code only the primary event or behaviour that led to recovery, with limited mention of secondary events.

JL1 responses across three formats

GL1 and JL1 think aloud informants' codings were discussed in chapter 6, and some aspects of GL1 AC responses were discussed above. For comparison purposes across the three reporting formats of think aloud, NCR (for which no GL1 data exists) I have chosen to focus on data from JL1 informants only. Table 7.25 (at the end of this chapter) summarises the most frequent processing operations encountered in JL1 informants' think aloud (TA) and NCR protocols, and on AC informants' manuscripts. Because the number of JL1 think aloud, NCR, and AC informants differed (*n*24; *n*=10*; and n*=24, respectively) the table shows the percentages of each group who entered the listed codings for each deletion. For some deletions the totals add up to more than 100%, and these represents contexts in which more than one processing behaviour was judged to have played a role in filling the blank. Thus, in deletion (3), seven out of ten NCR informants' protocols were coded as indicating use of 'same sentence forward' information, i.e. information contained in the sentence following the blank ('…international') while six out of ten mentioned, or also mentioned, information

preceding the blank, or 'same sentence backwards' ('...it's local character'.) One informant appeared to have relied on phrasal or collocational knowledge:

"[...] I PUT national event.. LIKE *Koshien* [the all-Japan high schools baseball tournament]."

The more commonly entered codes are given first, but a coding seen as representing partial success is listed only where it was intuited to represent a level of understanding of the nature of an (at that moment) *un*recovered item. In other words, where a verbal report informant recovered a SEMAC filler such as deletion (12) 'came', only then to say that "..it must be a verb anyway" this was not counted as an instance of word class selection. Examples of authentic 'WC' coding would be (1) where the informant realised that a verb was needed in, say, item (23), and subsequently filled the blank, or (2) could not fill blank (23) appropriately, but could state that a verb was needed. '???' in think aloud and NCR denotes no coding possible, and in AC no coding made by the informant. Grey-shading is used to show items which appear to have been processed together by the majority of informants. Once again I emphasise that some think aloud events could not be clearly allocated on the basis of concurrent or retrospective informant data, so that intuition played some role in tabulating these responses. NCR events were almost all fairly readily codeable, and the problem of allocation did not arise with AC responses.

Tabulation of all JL1 codings reveals inequalities in the extent to which the three reporting formats elicited second or subsequent 'recovery' codings (i.e. other than

'???') to accompany the most frequent one. Given the considerable variations involved, this aspect is perhaps best expressed by giving percentages of informant/item interactions in which this occurred in each format. These are shown below.

Think aloud	NCR	AC
16.5%	7.2%	4.4%

Figure 7.24: Percentage frequency of secondary event codings

It must be borne in mind that certain events ('SOUnding out', etc.; cf. chapter 6) were excluded from the tally here on the grounds that they are assumed not to be format-independent. The fact that these were most likely to be found in think aloud data has reduced the percentage shown above for that format. In other words, think aloud elicits secondary events more frequently (based on my best estimate of salience to recovery, on the order of 2.5%) than the figure suggests. Even allowing a margin of error through misinterpretaton or misallocation of events on my part, it seems clear that think aloud data is potentially the richest in terms of processing 'depth' recorded.

In just under 85% of items, the same processing event appeared as most frequent across all three formats. This seems to support the notion that cloze items typically are typically recovered by one best or most obvious route or cue-uptake. As no clear association was noted between cloze success and the appearance in the data of secondary events, this single most salient behaviour seems to be

adequate as a route to recovery. Differences in coding across formats do occur, although at least some of these may reflect differences in informant proficiency and/or knowledge. One illustration may be seen in item (21), in which extratextual knowledge ('KOW') played the dominant role in recovery in all three formats. In think aloud, however, 55% of informants indicated (or also indicated) use of cues from an earlier paragraph ('EP') but in NCR (as far as we can gather from their verbal reports) and AC all who filled the item relied solely on prior knowledge. Again, we may infer from this that the cues in an earlier paragraph played only a confirmatory or supportive role. In item (23), we see a similar event in the use by 21% of think aloud informants of translation. As this coding does not appear in the data from NCR or AC informants we must assume that it was not salient enough to meet the informants' 'criterion' for reporting.

7.24 Conclusion

In this chapter I have described a trial of questionnaire items as a means of gathering insight into cloze processing, particularly of aspects that are less well reflected in verbal report. Questionnaires were found to be a useful supplement to other data-gathering procedures, provided ways can be found to ensure that exposure to questions does not bias the informant's subsequent processing behaviour. Rather than administer questionnaires post-task, as in Nevo 1989, it was suggested that questions be set individually, as close as possible to the processing event in focus.

The development was then outlined of a procedure in which cloze test-takers categorise that subset of their own processing behaviours which they (a) *are able* and (b) *choose* to report. I have rationalised the use of such a reporting format in terms of its efficiency, but have also discussed limitations in the spectrum of the data gathered through AC in relation to that that obtainable via the verbal report procedures discussed in chapter 6. I have also outlined some aspects of the AC task that may to some extent compensate for these limitations. The psychological reality to informants of the categories used was assessed and found satisfactory, and some insight was gained into their assessment of categories' usefulness or salience. Further, it was established that informants' coding of their behavior appeared to be stable across two comparable cloze passages. Overall, as with verbal report, the picture of AC informants' processing lends weight to the conception of natural cloze as calling for processing at a predominantly local level.

The overall productivity of the three 'routes' to the categorization of processing behaviour, think aloud, NCR, and AC, was compared in terms of comprehensiveness and depth of coverage. NCR was found to most reliably provide codeable data about recoveries, followed by AC and then think aloud. Although the procedures fairly consistently recorded the same recovery events as most frequent (and thus hypothesised to be most salient in recovery) for each item, it was noted that AC tended to record fewer secondary or subsequent codes than either verbal report format. This, it was suggested, may be seen as either a

shortcoming of the AC procedure or, as with NCR, a reflection of the fact that informants themselves choose which events to report or code, hence allowing them to select only the behaviours or cues that played a genuine role in recovery. Pros and cons of this shifting of the burden of categorizing processing events onto the informant herself were discussed in section 7.8, but I suggest again that she may be the best judge of which behaviours were salient in her recovery of a cloze deletion, and which merely co-occurred.

Deletion	TA				NCR				AC			
1 four	KOW				KOW				KOW			
	100				100				100			
2 lost	Phr/Col	WC	???		Phr/Col	WC	???		WC	Phr/Col		
	67	42	33		70	20	20		71	50		
3 national	SSf	SSb	Phr/Col	???	SSf	SSb	Phr/Col		SSf	SSb	Phr/Col	???
	62	46	29	12	70	60	10		55	12	17	8
4 against	Phr/Col	???			Phr/Col	Gra	???		Phr/Col	???		
	79	21			80	10	10		91	8		
5 No	Phr?col	???			Phr/Col	???			Phr/Col	???		
	79	21			90	10			83	16		
6 the	Gra	???	SPb		Gra	SPb			Gra	???		
	71	17	12		90	10			91	8		
7 records	Phr/Col	???	WC	KOW	Phr/Col	WC	KOW		Phr/Col	WC	KOW	
	71	29	25	17	80	30	20		75	21	17	
8 took	Phr/Col	Gra	???		Phr/Col	Gra			Phr/Col	Gra	WC	???
	62	25	12		80	20			71	12	12	8
9 below	Gra	WC	KOW/Log	???	Gra	WC	Phr/Col	Gra	Gra	Phr/Col	WC	
	54	62	12	8	40	40	33	10	54	42	37	
10 gathered	Phr/Col	WC	???		Phr/Col	Gra			Phr/Col	Gra		
	71	33	12		91	9			83	17		
11 no	Gra	???	KOW		Gra				Gra	???		
	71	29	8		100				91	9		
12 a	Gra	???	KOW		Gra				Gra			
	71	29	8		100				100			
13 were	Gra	???			Gra	???			Gra			
	83	17			90	10			100			
14 order	Phr/Col	Tr	WC	???	Phr/Col	WC			Phr/Col	WC	???	
	71	55	21	12	70	50			87	17	12	
15 included	SSf/Log	WC	Tr	???	SSf/Log	WC	Phr/Col	???	Phr/Col	Log	WC	???
	42	33	25	17	60	30	20	20	50	33	8	8
∫16 such	Phr/Col	???			Phr/Col				Phr/Col	Gr		
	96	4			100				87	12		
17 the	Gra	???			Gra	???			Gra	???		
	79	21			90	10			87	12		
18 also	EP	Gra	???	KOW	EP	SSf			Gra	EP	SSf	???
	55	29	21	8	80	50			58	17	12	12

19 of	Gra	Tr	WC	???		Gra	Phr/Col	WC		Gra	Phr/Col	???								
	66	29	17	4		50	50	10		46	37	17								
20 the	Gra	???	KOW			Gra	EP			Gra	???									
	79	21	4			80	20			87	4									
21 jumping	KOW	EP	SPf/LP			KOW	Nec?			KOW	???									
	62	55	12			100	10			96	4									
22 The	Gra	???				Gra				Gra	???									
	62	37				100				92	8									
23 dedicated	Phr/Col	WC	Tr			Phr/Col	WC	Gra		Phr/Col	Gra	WC								
	71	33	21			90	30	10		79	21	4								
24 of	Phr/Col	Gra	???	WC		Phr/Col	WC	Gra		Phr/Col	Gra	???								
	67	17	21	12		80	20	10		67	25	12								
25 that	Tr	Phr/Col	SSb	Gra	KOW	KOW	SSb	Gra		Phr/Col	KOW	SSb	Gra							
	55	29	25	12	8	50	30	20		37	33	21	8							
26 aside	Phr/Col	WC	???	Nec?		Phr/Col	WC	Nec?	???	Phr/Col	WC	Nec?	???							
	79	33	12	8		90	20	10	10	96	4	4	4							
27 sixth	SPb/Log	KOW	???			SPb/Log				SPb	Log									
	91	4	4			100				62	37									
28 were	Gra	???				Gra				Gra										
	87	12				100				100										
29 olive	KOW	Phr/Col				KOW	Phr/Col			KOW	Phr/Col									
	62	37				90	10			67	29									
30 was	Gra	???				Gra				Gra	???									
	83	17				100				96	4									
31 the	Gra	???				Gra	???			Gra	???									
	79	21				80	20			87	12									
32 the	Gra	???				Gra	???													
	75	25				90	10													

Figure 7.23: Codings applied to JL1 think aloud & NCR protocols, and to AC manuscripts

CHAPTER 8: CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

8.0 Introduction

In this chapter I sum up some of the conclusions I have drawn from my experience of gathering verbal report and other data about the picture we may draw of how cloze task-takers go about the task. I look first at the advantages and drawbacks of think aloud, NCR and AC and the implications of these for the kinds of data that can be elicited. I then sum up what the data can tell us about how cloze tasks are carried out. Finally, I offer some suggestions for further research employing variants of the procedures discussed in earlier chapters.

8.1 Some conclusions about data-elicitation procedures

Even without taking into account those informants who seemed to find the think aloud task beyond them, it will be clear from the protocol data shown in chapter 6 and Appendix 5 that think aloud informants can vary widely in terms of how extensively they report, how interpretable their verbalizations are, and how much these tell us about what was going in their minds during the task session. Training in thinking aloud has been used by some researchers, although (cf. Pressley & Afflerbach 1995; Cohen 1997) this is by no means always described with any precision. Despite the force of their arguments in the original 1984 edition of their book, Ericsson & Simon 1993 appear to have had second thoughts on this question, and some forms of informant training may now have their imprimatur. 'Directive' training in think aloud reporting was avoided in my own data-elicitation for the reasons given in chapter 4, but the question of whether training actually boosts informants' overall productivity is anyway far from settled. Neumann 1995, for one, suspects that it may rather have a homogenizing

effect on the content of their verbal reports, and suggests that the fact of training may seriously compromise the comparison of collected data with that gathered from untrained or differently trained informants.

Despite its sometimes limited interpretability, and the high time demands it places on the researcher, think aloud *potentially* provides the richest and deepest data we can currently hope to glean from informants. NCR, the variant form of think aloud discussed in chapter 6, appears to preserve only part of the spectrum of data which conventional think aloud may record, but my suggestion is that the opportunity-cum-need to 'edit' reports allows or encourages NCR informants to select for report the events and cues which made a real contribution to their cloze recoveries. As it is by no means always possible for a third party to the informant-text interaction to distinguish between processing events which were salient to recovery and those which were incidental (Think aloud informants are not always able to do so clearly either, when interviewed post-task.) the informant's immediately retrospective intuitions expressed through NCR report may provide the most reliable assessment of this. Although it is not clear how much 'thinking space' the cognitive load of NCR leaves free, it appears to be more practical than in think aloud to set informants additional tasks such as rating individual items on a scale of difficulty: such additional tasks are carried out more consistently in the NCR task format.

As we saw in chapter 6, however, NCR does 'lose' data at least potentially available in think aloud such as the strict sequence of processing events (and quite

possibly the triggers for certain events), hesitations, and 'false starts' or superceded misinterpretations. If these are central objects of study then think aloud must be the preferred reporting format, whereas if it is enough to know which events the informant herself thought salient enough in her processing to merit reporting, then NCR is definitely the more economical choice. As mentioned earlier, NCR requires less time to conduct, and less time to transcribe and, if necessary, to translate. The content of Ryou's protocol in chapter 6 was in fact fairly typical of NCR in terms of its depth and clarity of reporting, although other informants did use their L1 rather more extensively. NCR, then, may be a good choice of initial reporting format, with recourse to concurrent think aloud should NCR prove inadequate as an elicitation tool..

The question of language of reporting (LoR) was discussed only briefly in chapter 4. As Cohen 1998 points out, the balance of TL and L1 use in processing is not necessarily the same as that in reporting, and informants may be unable, post-task, to give more than the most general indication of which language dominated their processing at any given point. (Absent this information, it seems reasonable to assume that the L1 predominated.) None of my informants to date, even the most fluent in the TL, has gone entirely without some use of her L1 in think aloud or NCR reporting, or in post-task retrospection/response to questions; I take this as sufficient evidence for the position that informants must be allowed an *authentic* choice of LoR. This may require the researcher to demonstrate that she can understand informants' verbal data well enough to interpret it accurately. The

majority of both my GL1 and JL1 informants have shown a genuine concern for confidentiality/anonymity, and it may be that for some of the latter their perceptions of my ability to autonomously interpret their L1 constrained what they felt able to report.

Orientations to tasks were described in some depth, as many writers (Cohen 1998; Afflerbach 2000) on verbal report as a research methodology require. In this area a drawback of AC procedure must be noted: while neither think aloud nor NCR requires the informant to 'learn' anything beforehand, AC requires orientation and practice before any data can be gathered. (The fact that a number of AC informants attended more than one orientation session, cf. chapter 7, suggests that I may have underestimated the need for preparation.) This initial outlay of time and effort must be set against the substantial savings in both that follow from the absence of any need to transcribe data. As may be seen in the detailed item comparisons presented in Appendix 4, and as discussed in chapter 7, AC informants appear to code only the most salient or dominant events, so that this procedure, too, loses some of the potentially captured by verbal report, and indeed some of that preserved in NCR protocols. Against this, AC does appear to leave room for attention to sub-tasks other than selecting codes, and ratings of item difficulty, for example, can fairly reliably be gathered. 'Marking up' of AC manuscripts can also provide a more precise indication of which passage content, if any, was translated into the informant's L1 than can typically be gleaned from think aloud data. Moreover, graphic links can trace connections between fillers,

passage items, or spans that existed in the informant's mind during processing. Finally, written comments on the manuscript can help clarify or expand on what coding and marking-up does not convey. As with the choice between think aloud and NCR, the decision to gather data via a checklist approach such as AC is likely to depend on what, or how much, one hopes to find out, and from how many. Another drawback of AC coding which must be recognised is the limitation of triangulatory evidence for coding decisions to any graphic or written information inscribed on the manuscript. The amount and usefulness of this supplementary/confirmatory information may vary considerably among informants, however, and further study is needed into whether and by how much additional practice can raise informant productivity in this area.

Another area of concern must be to gather data from the largest possible number of informants. The data presented here unavoidably contains only some of the interesting (but unrepresentative) idiosyncratic processing behaviours and interpretations of passage content I have noted, but there appears to be recognition (Pressley & Afflerbach 1995) that gathering think aloud data from small informants samples risks bias from idiosyncratic behaviour.

Whichever tasks or task formats we use in gathering data about task-processing, a number of issues are clearly identifiable. First, as Cohen 1998; Afflerbach 2000; Boren & Ramey 2000 and other writers in the field have argued, we need to ensure that data are as comparable as they can be. This calls for reasonably detailed elucidation of coding categories, and illustration of their application to at

least some informant data, as well as descriptions of any task-modeling and/or training given to informants pre-task. This is not to suggest that different studies should employ the same techniques, or even represent data in very similar ways, but merely to echo the view that one cannot see the good and bad points of a data-gathering procedure unless one knows in some detail how it was carried out, and why in that way. I would echo, too, the suggestion in Boren & Ramey 2000 that it is questionable to cite in justification of one's methodology authorities like Ericsson & Simon 1984/1993, but then to collect data in ways which may not meet their criteria. I have argued in chapters 4 and 5 that the Ericsson & Simon model may not adequately describe some aspects of the think aloud data informants generate while engaged on language tasks, and wonder aloud again about whether the model was ever really intended to be used with extended tasks of this kind in which recall of prior knowledge, judgement about performance, etc. may be unavoidable.

8.2 What can we conclude about how cloze tasks are carried out?

Looked at broadly, the data discussed in previous chapters suggests that cloze task-taking is at once simpler and more complicated than many users of the procedure might expect. As codings of protocol data gathered via the two verbal report formats and by AC show, in the majority of cloze items one processing operation appears to play a dominant role in recovery. The two most frequent event categories in the data are 'Gra(mmatical knowledge' and 'Phr(asal)/Col(locational/ Idi(omatic) knowledge', clearly supporting the claim of

Alderson 2000, Klein-Braley 1981 and others that natural cloze predominantly taps local-level constraint. (This holds true even though informants appear at times to have selected inappropriately between these two categories.) The comparative rarity of instances of long distance cue uptake such as 'E(arlier) P(aragraph)' and 'L(ater) P(aragraph)' in filling blanks appears to be at odds with the repeated claims of proponents of cloze (cf. Oller & Jonz 1994) that it also targets processing of text at a global level, but in mitigation it must be pointed out that texts vary in the amount of long-distance 'cross-referencing' they permit or require (Halliday & Hasan 1976) and in how this is manifested. The stimulus task passage used here, OLYMPICS, may in fact demand comparatively little in the way of long-distance cue activation. The processing demands of the passage, then, will affect the range of processing events required. A passage which requires only a limited range of processing behaviours may thus have a homogenizing effect on informants' reports, and this effect may be amplified to some extent by a procedure that tracks only the most salient events. Natural cloze passages, however, are typically generated from real-world source texts, rather than written to order to elicit particular behaviours. That said, Alderson 2000 has suggested that gap-filling, or rational cloze, tasks should be preferred over natural cloze, on the grounds that the former allows the test constructor to better target aspects of language processing; to some extent selective deletion can get around the possibly limited processing requirements set by real-world texts plus mechanical reduction of redundancy.

A further important category in cloze recovery is 'Knowledge Of the World', as this makes up in power what it may lack in frequency of application. It is hard to see patterns in the application of KOW in the passages used here, except insofar as (1) comparatively few deletions called for extratextual information, and (2) informants appeared to apply whatever prior knowledge they possessed when this was called for. Think aloud protocols suggest that application of extratextual information may be less a kind of 'fallback' behaviour, following an unsuccessful search for cotextual cues, but rather typically the first event identifiable as present in the span of verbalization about items to which it is pertinent. Unsurprisingly, possession of relevant extratextual knowledge typically leads to successful filling of a blank, so that 'KOW' codings are associated with slightly higher SEMAC scores.

'KOW' can also be misapplied, however, such as in OLYMPICS deletion (18), which some informants filled with the name of a sporting event rather than the SEMAC 'also'. Prior knowledge, then, may be double-edged. (Post-task interviews confirmed that some of these informants had failed to note, or realise the significance of, the absence of a comma before 'boxing', while others seemed to have inferred that the listing of events that precedes and follows this deletion required another event still.) This confusion points up the difficulty in coding application of extratextual knowledge, for those who mistakenly inferred that another event was needed in (18) could also be said to have been referring to cues in an earlier paragraph ('EP') or information later in the sentence ('SSf'), or even

making a logical inference ('Log'.) The fairly common co-occurrence of 'KOW' and 'Log' in think aloud protocols reflects the difficulty of separating these. Logical inferences often appear to be triggered by extratextual knowledge, which in turn may be verbally expressed in terms of 'this must be/cannot be *x*'.

Unambiguous associations between other events and cloze success are, as discussed in chapter 6, even less common. The roles of translation and L1 item glossing in cloze success are particularly unclear, for it is intuitively plausible (and I experience this myself on a daily basis) that other cognitive events may be going on at the same time as passage content is being translated. Translation may begin before a filler has been selected and then continue beyond that point. It may thus be unsafe to assume that, when a filler emerges following a translation event, recovery was necessarily due to that rendering into the L1. Unless supplementary information is available, however, 'Tr' may be the only realistic coding option in such instances.

Such information is sometimes available in the form of post-task interview comments, but these are not reliably insightful. JL1 pair think aloud informant Mitsuo's retrospections about his processing of OLYMPICS deletion (24) '..heroes <u>of</u> the day' is a case in point. Mitsuo claimed to have been known and "thought about" the TL chunk '*of the day*', but that part of his protocol is almost entirely in the L1 and provides no evidence of this. When asked to do so, however, Mitsuo was able to quickly and accurately cite some TL uses of the structure: '*hero of the hour*'; '*hero* of the school', so that his claim to have applied TL

phrasal knowledge at some level cannot be discounted.

In talking of the 'predictability' of cloze I refer to the fairly good 'fit' between the predictions of German and Japanese L1 consultants (almost all educators of one kind or another) whom I asked to suggest the kinds of information task-takers from the same language backgrounds would be likely to possess and to use (grammatical knowledge used here, informants likely to know the necessary collocation there, over there knowledge of the topic is required, etc.) and those subsequently found to be dominant in those deletions. A German or Japanese consultant might thus note that close L1 parallel structures existed to, say, deletion (24), above, but she could not be expected predict whether an informant would access the L1 parallel as a salient cue to recovery, let alone make this explicit in verbal report.

This 'commonsense' predictability of cloze recovery operations applies to the question of how an informant *might* be able to fill a blank rather than how she will do so. In a country like Japan, which was until fairly recently a 'stimulus poor' environment for foreign language acquisition outwith the education system, the highly centralised school English syllabus allowed educator- consultants to predict fairly accurately which TL vocabulary and structures others would have been exposed to, and even how extensive that exposure was likely to have been. With the rapid spread of English-language media in Japan, and the relaxation of school foreign-language curricula, this predictability is likely to diminish.

All in all, the conclusions which I feel can safely be drawn from the data do not extend much beyond the above, for the difficulty of assessing the role in cloze recovery of 'secondary' events identified in think aloud verbal report data is considerable. Where, and this was a common event, an informant is herself unsure post-task about the salience of a given event in filling a deletion, it may be unwise to try to make the decision for her.

Although think aloud may potentially provide richer data than other procedures, this is only of value when fairly secure interpretation is possible. It may be, as I suggested above, better to leave to informants themselves the task of deciding what did, or did not, play a role in their recovery of a deleted cloze item.

8.3 Possible avenues of future research

Although all of the data reported here was gathered from traditional cloze, there would be good reason to apply think aloud, NCR or AC to what Alderson calls 'gap-filling tasks', or 'selective-deletion' cloze, for the popularity of this task format is growing (see chapter 2) as test-constructors seek to achieve more control over which passage items are targeted. Although natural cloze, to use J.D. Brown's label, is probably still more common (apparently undergoing periodic revival in the Asia-Pacific region) selective-deletion cloze/gap-filling task formats seem set to overtake it in the not-too-distant future.

Gap-filling tasks (see below) might be even better suited to think aloud. With additions to cover cues from morphological form, the codeset used in this paper could be applied to the analysis of lexical inferencing tasks, which in my own

view are a step closer to the kind of reading comprehension task that students face in lessons, as well as in test situations.

Since, as think aloud data and AC underlining and graphic marking-up have indicated, the difficulty of a cloze task does not lie solely in the blanks, a given cloze passage may to some extent already incorporate lexical inferencing as part of its challenge. It might be productive to combine the two meaning recovery tasks in more controlled ways, such as by deleting half of a target set of high-challenge passage items and looking at how closely informants' recoveries of word of items deleted and items still extant resemble each other.

Finally, although (see chapter 2) ideas of 'coherence' and 'discourse' cloze may not yet be fully worked out, there is no debate (Nattinger & DeCarrico 1992) that some multi-word 'chunks' or 'routines' carry special meaning or perform particular roles in creating 'textuality'. The processes by which informants recover such units could also be usefully addressed by verbal report in one form or another.

I discussed in chapter 7 some of the individual 'palettes' of task-processing behaviour codes which informants had assembled. It would be interesting to see whether, and if so how, these self-selected sets evolve over time. One strength I have noted of the AC format is that once the codeset has been 'acquired' only a brief refresher session should be needed for it to be successfully re-used in follow-up data-elicitation sessions. It would thus be possible to invite experienced

informants to take a further AC cloze (or other) task some months later in their language learning careers. This is possible even if the informant and researcher are still in contact but no longer in proximity: AC requires no facilities beyond a desk and a pencil, and can be administered by mail or even email.

Dispensing with the researcher's presence

Observation of informants at work may provide useful data on aspects of the task that typically go unrecorded (in the absence of video) in think aloud, such as 'observable movements' within the passage (searching ahead of the item; looking back at previous text, etc.) but this is really only practical with individual informants or very small groups. and the trade-off in more drawn-out data-elicitation may not be deemed worthwhile. My observation notes of solo-condition think aloud and NCR sessions suggest that in many cases little information was gained through observation that was not also available on the recording (although, as noted in chapters 5 and 6, in some instances useful insights were gleaned through direct observation) and the researcher may find that she is acting largely as timekeeper, reminer-to-verbalise, and LL technician.

Once a verbal report procedure's kinks have been ironed out, however, it may be possible to dispense entirely with the researcher's presence during reporting. With the advent of extremely portable and increasingly inexpensive digital recorders, there is no reason why think aloud should not become a kind of 'take-away task' to be done at an informant's own convenience. She might even (equipment and motivation allowing) make more than 'pass' at the task, adding or elaborating on

information until she is happy with the picture it provides of her processing. This proposal would of course do nothing to ease the burden of transcribing the resulting protocol, but it could allow the gathering of data from a large number of informants within a short time frame even in the absence of LL facilities. Opportunities to query points authentically post-task would decline, of course, as the informants' recordings would have to be audited before questions could be formulated.

Another possibility perhaps worth exploring further is that of asking informants to carry out a cloze (or similar) task and at each recovery to report in writing how they arrived at their chosen filler. A recent small-scale trial of this procedure with a group of eight student teachers produced protocols remarkably similar to transcribed NCR verbal reports, with some including events by no means reliably encountered in verbal reports of item processing : questioning of the need for a filler in a given blank; estimates of confidence in the chosen filler; mention of alternatives to it, etc.

This task was done in the informants' own time, and although not all recorded how long they spent on it, none reported that she had found the task excessively demanding. The fact that they had a written record to refer to, two informants later mentioned, allowed them to take a break part-way through the task, and the fairly numerous additions or elaborations to written comments (plus a few revisions) were clearly identifiable through carets, arrows, deletions and rewriting. It should be noted, however, that in spite of a reputation for vagueness of meaning (cf.

Maynard 1997) Japanese is a particularly 'condensed' and economical written code: on users of alphabetic scripts a written reporting task might place a rather greater time-demand. One potentially productive variation on the 'free' written protocol combines AC-style coding decisions with written elaboration, though (unlike AC as used to date) emphasizing the primacy of the *written* material and with codes or descriptors as a kind of shorthand for unambiguous events such as uptake of grammatical cues, application of extratextual knowledge, etc.

Still another verbal reporting format?

As already discussed, a majority of my solo-condition think aloud informants claim to have found the task fatiguing and stressful, and both NCR and AC were seen as ways of lightening the informant's burden as well as that of the researcher. Other variations on the theme of think aloud may be useful, however. One variant of think aloud which I have recently begun to explore requires informants to report on their processing of only a few cloze items within the task, but in as much detail as possible. This level of effort cannot, I think, reasonably be expected for each item in a natural cloze within a reasonable task-session, so that it would be necessary to allot different items to different informants. The data this reporting format generates may be very rich:

" [...] WELL.. came from all parts of Greece IS OKAY HERE... BUT MAYBE travelled IS BETTER BECAUSE.. I THINK THE IDEA IS THAT THEY CAME A LONG WAY.. AND IN THOSE DAYS TRAVEL WASN'T SO EASY.. SO I'LL PUT traveled.. but.. even as a spectator.. OKAY no married woman was admitted even as a spectator.. I LEARNED THAT [grammatical structure in school].. no something something even something..OKAY [deletion] TWELVE IS *a* spectator BECAUSE IT HAS TO BE SINGULAR [...]"

(OL/Aset/Ke)

Interestingly–and encouragingly insofar as it suggests that very much of the informant's interaction with the text is open to report–an additional coding for 'author's perceived intent' would be needed to describe the extract above. Such a technique may be especially useful in looking at the processing of 'rational' or 'selective-deletion' cloze, or gap-filling tasks, which are typically lower in item-density than 'natural' cloze. Gap-filling tasks, according to Alderson 2000, are to be constructed according to given criteria or hypotheses rather than the broad notion of linguistic redundancy, and the fit between processing events hypothesised in a given context, and those actually occurring, may be open to examination via informant self-reports of one kind or another. \

Informants' task-format preferences

Last but not least, my experience of administering verbal report tasks–in particular concurrent think aloud–has led me to the view that in order to be gather data ethically we must take into account informants' affective responses to the task. By this I mean not simply minimizing informants' fatigue and/or stress, but rather by providing them with the tools to perform the task as far as possible to their own satisfaction. Post-task interviews and informal discussion with individual GL1 and JL1 informants suggested that some had a strong sense of how well or otherwise they had managed, in their own estimation, to give a picture of their on-task behaviour. Some informants in fact expressed frustration on auditing their own protocols post-task, noting that they had left their recovery of some items insufficiently explained. The suggestion was even made that I should have more

rigorously reminded students to verbalise.

Given that think aloud informants have been known to return spontaneously to an earlier deletion, or to make more than one 'pass' at the task as a whole, it is hard to see how more consistent or fuller reporting could be achieved except by obligating them to report at given points (see Appendix 6: 'Encouraging verbalization'.) It may be an advantage of the NCR variant of verbal reporting that, as reported in chapter 6, fewer recoveries seem to go unreported. This may in part be due to that format's enforced 'suspension' of processing in order to report recoveries as they happen.

In an ideal humanistic data-gathering project, informants might be orientated to a variety of reporting formats and allowed to choose those which best suit their perceived abilities and preferences. How realistic such an arrangement might be is open to question, but it should be noted that allowing informants a choice of format would *not* inevitably lead to everyone opting for pair-condition think aloud—popular as that format proved to be among my own informants. In a recent classroom 'opinion survey' I outlined to a group of 22 students the data-elicitation tasks discussed in this paper—think aloud in solo- and pair-condition, NCR , and finally AC—and asked them to imagine themselves as prospective informants and to rank the four task formats in their personal order of preference. (Questions about the formats were invited and answered in detail.) Just over half of the group marked pair-condition think aloud as their first choice, while all but a few placed NCR either first or second. AC also placed just ahead of solo-condition

thinkaloud in ranked choices. Even though pair-condition reporting was the most popular reporting option, there appear to be plenty of informants who prefer to report alone, although by no means all would choose concurrenmt think aloud.

A glance at the contents page of almost any recent 'introductory' text on language testing reveals that ethical issues are increasingly central to the field: test-takers expect to be given tests for which they are prepared and which they feel will provide an accurate picture of their ability. I would argue that the same expectations may apply even in pseudo-tests, and that we cannot assume that the distinction between an actual test and a data-gathering, test-like task is very sharply drawn in the minds of those from whom we get our data. For some of my informants at least, it appeared that a cloze was a cloze was a test.

Bob Gibson

Wadamachi, July, 2005

i

Bibliography

Aborn, M., Rubenstein, H. and Sterling, T., 1959, Sources of Contextual Constraint Upon Words in Sentences, Journal of Experimental Psychology 57: 171–180

Addison, A. and Vogel, K. (Eds), 1987, Lehren und Lernen von Fremdsprachen im Studium, Bochum: AKS-Verlag

Afflerbach, P., 1990, The Influence of Prior Knowledge on Expert Readers' Main Idea Construction Strategies. Reading Research Quarterly, 25: 31-46

Afflerbach, P. ,2000, Verbal Reports and Protocol Analysis (In Kamil, M.. et al, 2000, A Handbook of Reading Research Vol. III, Newark NJ:National Reading Association)

Aitken, K., 1977, Using Cloze Procedure as an Overall Language Proficiency Test, TESSOL Quarterly 11,1:59–67

Aizawa, K., 1998, Lexical Inferencing Cues in Reading and Japanese Learners of English, Daigaku Eigo Kyouiku Gakkai Bulletin 11, 1998: 1-19

Alderson, J.C., 1978, A Study of the Cloze Procedure with Native and Non-Native Speakers of English, Unpublished Phd Thesis, University of Edinburgh

Alderson, J.C., 1979a, The Cloze Procedure and Proficiency in English as a Foreign Language, TESOL Quarterly 13:219-27

Alderson, J.C., 1979b, The Effect on the Cloze Test of Changes in the Deletion Frequency, Journal of Research in Reading 2:108-119

Alderson, J.C., 1983, 'The Cloze Procedure and Proficiency in English as a Foreign Language'(In Oller, J.W.Jr., 1983)

Alderson, J.C., 1990, Testing Reading Comprehension Skills (Part II), Reading in a Foreign Language 7,1:465-503

Alderson. J.C., Clapham, C. and Wall, D., 1995, Language Test Construction and Evaluation, Cambridge: CUP

Alderson, J.C. and Urquhart, A.H., 1984, Reading in a Foreign Language, London: Longman

Alderson, J.C. and Urquhart, A.H., 1985, The Effect of Students' Academic Discipline on Their Performance on ESP Reading Tests, Language Testing 2,2:194–204

Alderson, J.C., 2000, Assessing Reading, Cambridge:CUP

Allan, A., 1992a, The Influence of Text Characteristics, and Test Item Format and Category on the Reading Strategies of ESL Students Attending a Tertiary Level Institution in Hong Kong, City Polytechnic of Hong Kong Research Report No.16, June 1992

Allan, A., 1992b, EFL Reading Comprehension Test Validation: Investigating Aspects of Process Approaches, Unpublished Phd Thesis, Lancaster University

Allan, A., 1995, Begging the Questionnaire: Instrument Effect in Readers' Response to a Self-Report Checklist, Language Testing 12,2:133–156

Ames, W., 1966, The Development of a Classification Scheme of Contextual Aids, Reading Research Quarterly 2,1:57–82

Anderson, J., 1976, Psycholinguistic Experiments in Foreign Language Testing, University of Queensland Press

Arntz, R., (Ed), 1988, Textlinguistik und Fachsprache: Akten des Internationalen Übersetzungswissenschaftlichen AILA-Symposions, Hildesheim, April 1987. Hildesheim: Olms

Bachman, L., 1982, The Trait Structure of Cloze Test Scores, TESOL Quarterly 16,1:61–69

Bachman, L., 1985, Performance on Cloze Tests with Fixed-Ratio and Rational Deletion, TESOL Quarterly, 19, 535-556

Bachman, L., 1990, Fundamental Considerations in Language Testing, Oxford: OUP

Baddeley, A., 1986, Working Memory, Oxford Psychology Series Vol.11, Oxford: OUP/Clarendon

Ballstaedt, S. and Mandl, H., 1984, Elaborations: Assessment and Analysis (In Mandl, Stein and Trabasso 1984:331-353)

Becker, J., 1972, An Information Processing Model of Intermediate-Level Cognition Cambridge, MA: Bolt Beranek and Newman Inc., 1972 (BBN Rep. No. 2335, Doctoral Dissertation)

Bensoussan, M., 1984, A Comparison of Cloze and Multiple-Choice Reading Comprehension Tests of English as a Foreign Language, Language Testing 1,1:101-104

Bensoussan, M., Goldenblatt, L. and Kreindler, I., 1984, Changing the Difficulty Level of Multiple-Choice EFL Reading Comprehension Questions. Language Testing, 1,1:105-114.

Bensoussan, M. and Laufer, B., 1984, Lexical Guessing in Context in EFL Reading Comprehension, Journal of Research in Reading, 7, 1, 15-32.

Bensoussan, M. and Ramraz, R., 1984, Testing EFL Reading Comprehension Using a Multiple-Choice Rational Cloze, Modern Language Journal 68:230-239

Bensoussan, M., 1990, Redundancy and the Cohesion Cloze, Journal of Research in Reading 13,1:18-37

Bernhardt, E., 1991, Reading Development in a Second Language: Theoretical, Empirical, and Classroom Perspectives. Norwood, NJ: Ablex Publishing Corp.

Bernhardt, E. and Mendez, C., 1989 Crosslinguistic Text Processing Strategies: Native Readers of English Reading in German (In Dechert and Raupach (Eds) 1989)

Bialystok, E., 1978, A Theoretical Model of Second Language Learning, Language Learning 28,1:69-83

Bialystok, E., 1983, Some Factors in the Selection and Implementation of Communication Strategies (In Faerch and Kasper (Eds) 1983)

Bialystok, E., 1984, Strategies in Interlanguage Learning and Performance (In Davies, Criper and Howatt (Eds) 1984)

Block, E., 1986, The Comprehension Strategies of Second Language Readers, TESOL Quarterly 20,3:463-494

Block, E., 1992, See How They Read: Comprehension Monitoring of L1 and L2 Readers. TESOL Quarterly, 26,2:319-343

Boren, M. and Ramey, J., 2000, Thinking Aloud: Reconciling Theory and Practice, IEEE Transactions On Professionsl Communications 43,3:261-277

Boring, E., 1953, A History of Introspection, Psychological Bulletin 50,3:169–189

Bormuth, J., 1965, Readability: A New Approach, Reading Research Quarterly 1:79–132

Bormuth, J., 1967, Comparable Cloze and Multiple-Choice Comprehension Test Scores, Journal of Reading 10:291–299

Börner, W., 1989, Planen and Problemlösen in Fremdsprachlichen Schreiben: Einige Empirische Befunde (In Klenk, U. , Körner, K., Thümmel, W. (Eds), 1989)

Börner, W. & Vogel, K. (Eds) 1995, Der Text im Fremdsprachenunterricht, Bochum: AKS.

Bransford, J.D. and Johnson, M.K., 1972, Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall, Journal of Verbal Learning and Verbal Behavior, 11: 717-726.

Brown, G., 1989, Making Sense: The Interaction of Linguistic Expression and Contextual Information, Applied Linguistics, 10,1:97-108

Brown, J.D., 1983, A Closer Look at Cloze Validity (in Oller, J.W. Jr. and Jonz, J. (Eds) 1994

Brown, J.D., 1989a, Tailored Cloze: Improved with Classical Item Analysis Techniques, Language Testing 6,1:19-31

Brown, J.D., 1989b, Cloze Item Difficulty, JALT Journal, 11:46-67, Tokyo: The Japan Association for Language Teaching

Brown, J.D., 1993, What are the Characteristics of Natural Cloze Tests? Language Testing, 10,2:93-116.

Burt, M.K. and Dulay, H.C. (Eds), 1975, New Directions in Second Language Learning, Teaching and Bilingual Education: Selected Papers From the 9th Annual TESOL Convention, Washington DC:TESOL

Butler, J., 1991, Cloze Procedures and Concordances: The Advantages of Discourse Level Authenticity in Testing Expectancy Grammar, System 19,1-2:29-37]

Carpenter, P. and Just, M., 1977, Integrative Processes in Comprehension (In Laberge, D. and Samuels, S. (Eds) 1977)

Carrell, P.L., 1987, Content and Formal Schemata in ESL Reading. TESOL Quarterly, 21,3:461-481

Carte, M., Petöfi, J. and Sözer, E., 1989, Text and Discourse Connectedness, Amsterdam: Benjamins

Carver, R., 1994, Percentage of Unknown Vocabulary Words in Text as a Function of the Relative Difficulty of the Text: Implications for Instruction, Journal of Reading Behaviour 26,4:413-437

Casanave, C., 1988, Comprehension Monitoring in ESL Reading: A Neglected Essential, TESOL Quarterly 22,2:283-302

Cavalcanti, M., 1987, Investigating EFL Reading Performance Through Pause Protocols, (In Faerch and Kasper (Eds) 1987)

Chapman, L.J. (Ed), 1981, The Reader and the Text, London: Heinemann

Charolles, M., 1989, Text Coherence and Text Interpretation Processing, (In Conte, M. et al, 1989)

Chavez-Oller et al, 1985, When are Cloze Items Sensitive to Constraints Across Sentences?, Language Learning 35,2:181-206

Cleary, C., 1988, The C-Test in English: Left-Hand Deletions. RELC Journal, 19,2:26-38.

Cohen, A., 1980a, Testing Language Ability in the Classroom, Rowley, MA: Newbury House

Cohen, A., 1980b, On Taking Language Tests: What the Students Report, Language Testing 1,1:70-81

Cohen, A, 1981, Introspecting About Second Language Learning, Centre For Applied Linguistics, Hebrew U., Jerusalem

Cohen, A., 1984, Studying Second Language Learning Strategies: How Do We Get The Information?, Applied Linguistics 5,2:105-112

Cohen, A., 1988, Strategies in Learning and Using a Second Language, London: Longman

Cohen, A., 1991, Feedback On Writing: The Use of Verbal Report, Studies in Second Language Acquisition 13:133-159

Cohen, A., 1995, In Which Language Do/Should Multilinguals Think? Language, Culture, and Curriculum, 8: 99-113

Cohen, A. and Aphek, B., 1981, Easifying Second Language Learning, Studies in Second Language Acquisition 3:221-236

Cohen, A. and Hosenfeld, C., 1981, Some Uses of Mentalistic Data in Second Language Research, Language Learning 31,2:285-313

Cohen, L. and Manion, L., 1994, Research Methods in Education (4th Ed.), London: Routledge

Cohen, L., Manion, L. and Morrison, K., 2000, Research Methods in Education (5th Ed.), London: Routledge

Conrad, R., 1962, Practice, Familiarity and Reading Rate for Words and Nonsense Syllables, Quarterly Journal of Experimental Psychology 14:71-76

Conte, M. et al, 1989, Text and Discourse Connectedness, Amsterdam:Benjamin

Cook, V., 1986, Experimental Approaches to Second Language Learning, Oxford: Pergamon

Cook, V., 1992, Evidence for Multicompetence. Language Learning, 42: 557-591

Culhane, T., Klein-Braley, C. and Stevenson, D.K. (Eds), 1983, Practice and Problems in Language Testing 7: *Proceedings of the Seventh International Language Testing Symposium of the IUS, Colchester*: University of Essex

Cziko, G.A., 1978, Differences in First- and Second-Language Reading: The Use of Syntactic, Semantic and Discourse Constraints, Canadian Modern Language Review, 34,3:473-489.

Daly, J. et al, 1989, Concurrent Cognitions During Conversations: Protocol Analysis as a Means of Exploring Conversations, Discourse Processes 12:227-244

Davies, A. , Criper, C. and Howatt, A. (Eds), 1984, Interlanguage: Proceedings of the Seminar in Honour of S. Pit Corder, Edinburgh:EUP

Davies, A., 1990, Principles of Language Testing, Oxford: Blackwell

Dechert H. and Raupach, M., 1989, Identification of Text-Type as a Strategic Device in L2 Comprehension (In Dechert, H., and Raupach, M. (Eds) 1989)

Dechert, H. and Raupach, M. (Eds), 1989, Transfer in Language Production. New Jersey: Ablex

Dechert, H. and Sandrock, U., 1986, Thinking Aloud Protocols: The Decomposition of Language Processing (In Cook, V. (Ed), 1986)

Deffner,G., 1987, Lautes Denken Als Methode Der Datenerhebung (In Arntz, R., (Ed) 1987)

De Sola Pool, I. (Ed), 1959, Trends in Content Analysis, Urbana, IL: University of Illinois Press

Deyes, A., 1984, Towards An Authentic Discourse Cloze, Applied Linguistics 5,2:128–137

Dingwall, S., Mann, S. and Katamba, F. (Eds), 1982, Methods and Problems in Doing Applied Linguistic Research, Lancaster: University of Lancaster Dept. of Linguistics and Modern English Language

Dixon, N., 1981, Preconscious Processing, Chichester: Wiley

Duncker, K., 1945, On Problem Solving, American Psychological Association (Psychological Monographs 58)

Ebbinghaus, H., 1897, Über Eine Neue Methode Zur Prüfung Geistiger Fähigkeiten Und Ihre Anwendung Bei Schulkindern, Zeitschrift für Angewandte Psychologie 13: 401-459

Ellis, R., 1986, Understanding Second Language Acquisition, Oxford: OUP

Ellis, R., 1994, The Study of Second Language Acquisition. Oxford: OUP

Enkvist, N. and Kohonen, V., 1977, Cloze Testing: Some Theoretical and Practical Aspects (In Zettersten, A. (Ed), 1977)

Ericsson, K. and Simon, H., 1980, Verbal Reports as Data, Psychological Review 87:215-251

Ericsson, K. and Simon, H., 1984/1993, Protocol Analysis: Verbal Reports as Data, Cambridge, Mass: MIT Press

Faerch, C., 1984, Strategies in Production and Reception–Some Empirical Evidence, (In Davies, A. , Criper, C. and Howatt, A. (Eds), 1984)

Faerch, C. and Kasper, G., 1983, Strategies in Interlanguage Communication, London: Longman

Faerch, C. and Kasper, G., 1987, From Product to Process: Introspective Methods in Second Language Research. (In Faerch, C. and Kasper, G. (Eds), 1987)

Faerch, C. and Kasper, G. (Eds), 1987, Introspection in Second Language Research, Clevedon: Multilingual Matters

Feldmann, U, and Stemmer,B., 1987, Thin____ Aloud A_____ Retrospective Da_____ in C-T_____ Taking: Diffe_____ Languages - Diffe_____ Learners - Sa_____ Approaches? (In Faerch, C. and Kasper, G. (Eds) 1987)

Figurel, J.A. (Ed), 1965, Reading and Inquiry, Newark, Del: International Reading Associates

Finch, G., 2000, Linguistics Terms and Concepts, London: Macmillan

Finn, P.J., 1978, Word Frequency, Information Theory, and Cloze Performance: A Transfer Feature Theory of Processing in Reading, Reading Research Quarterly, 13,4:508-53

Fischler, I. and Bloom, P.A., 1979, Automatic and Attentional Processes in the Effects of Sentence Context on Word Recognition, Journal of Verbal Learning and Verbal Behaviour 18:1-20

Flammer, A. and Kintsch, W. (Eds), 1982, Discourse Processing. Amsterdam: North Holland

Foster, K., 1976, Accessing the Mental Lexicon (In Wales, R. and Walker, E. (Eds), 1976)

Freedle, R.O. (Ed), 1979, New Directions in Discourse Processing, Norwood, N.J.: Ablex

Freedle, R.O. and Dunn, R. (Eds), 1987, Cognitive and Linguistic Analysis of Text Reference, Norwood, NJ: Ablex

Freud, S., 1914/1917, On the History of the Psycho-Analytic Movement, Nervous and Mental Disease Monograph Series No. 25 (Trans. Brill, J.) New York: Nervous and Mental Disease Pub. Co.

Gallant, R., 1965, Use of Cloze Tests as a Measure of Readability in the Primary Grades (In Figurel, J.A. (Ed), 1965)

Garrod, S. and Sanford, A., 1985, On the Real-Time Character of Interpretation During Reading, Language and Cognitive Processes 1,1:43-59

Gass, S. and Mackey, A., 2000, Stimulated Recall Methodology in Second Language Research, Mahwah, N.J.: Erlbaum

Gerloff, P., 1987, Identifying the Unit of Analysis in Translation (In Faerch, C. and Kasper, G., 1987)

Gibson, B., 2002, Talking at Length and Depth: Learning from Focus Group Discussions (In Johnson, K. and Golombek, P. (Eds), 2002)

Gibson, B., 1993, Introspective Data Collection and the Investigation of Learner Strategies, Unpublished MSc, Dissertation, University of Edinburgh

Gibson, E. and Levin, H., 1975, The Psychology of Reading, Cambridge, Mass: MIT Press

Gilhooly, K., 1986, Individual Differences in Thinking Aloud Performance, Current Pyschological Research and Reviews 5,4:328–334

Glaser, B.G. and Strauss, A.L., 1967, The Discovery of Grounded Theory Analysis, Mill Valley CA: Sociology Press

Goodman, K.S., 1967, Reading: A Psycholinguistic Guessing Game, Journal of the Reading Specialist 6:126-135

Gordon, C., 1987, The Effect of Testing Method on Achievement in Reading Comprehension Tests in English as a Foreign Language, Unpublished Master's Thesis, Tel-Aviv University

Grabe, W., 1991, Current Developments in Second Language Reading Research, TESOL Quarterly 25,3:375-406

Graesser, A., 1981, Prose Comprehension Beyond the Word, New York; Springer

Graesser, A. and Kreuz, R., 1993, A Theory of Inference Generation During Text Comprehension, Discourse Processes 15:145-160

Gravetter, F. and Wallnau, L., 2000, Statistics for the Behavioural Sciences (5th ed.) Belmont, CA: Wadsworth

Green, A., 1998, Verbal Protocol Analysis in Language Testing Research (Studies in language Testing 5), Cambridge: CUP

Greenwald, A. G., Klinger, M. R., & Schuh, E. S., 1995, Activation by Marginally Perceptible ("Subliminal") Stimuli: Dissociation of Unconscious from Conscious Cognition, Journal of Experimental Psychology: General, 124: 22-42

Groner, R., d'Ydewalle, G. and Parham, R. (Eds), 1990, From Eye to Mind: Information Acquisition in Perception, Search and Reading, Amsterdam: North Holland

Grotjahn, R., 1986, Test Validation and Cognitive Psychology: Some Methodological Considerations. Language Testing 3,2: 159-185.

Grotjahn, R., 1987a, Ist Der C-Test Ein Lesetest (In Addison, A. and Vogel, K. (Eds), 1987)

Grotjahn, R., 1987b, On the Methodological Basis of Introspective Measures (In Faerch, C. and Kasper, G. (Eds)m 1987)

Grotjahn, R., 1992, (Ed), Der C-Test: Theoretische Grundlagen Und Praktische Anwendungen. Vol 1, Bochum: Brockmeyer

Grotjahn, R., 1993, (Ed), Der C-Test: Theoretische Grundlagen Und Praktische Anwendungen. Vol 2, Bochum: Brockmeyer

Grotjahn, R., Klein-Braley, C. and Stevenson, D. (Eds), 1987, Taking Their Measure: The Validity and Validation of Language Tests, (Quantitative Linguistics Vol 34) Bochum: Brockmeyer

Grotjahn, R., Klein-Braley, C. and Raatz, U., 1992, C-Tests in Der Praktischen Anwending: Erfahrungen Beim Bundeswettbewerb Fremdsprachen (In Grotjahn, R., (Ed), 1992)

Grundin, H. et al, 1981, Cloze Procedure and Comprehension: An Exploratory Study Across Three Languages, Journal of Reading Research 4,2:104-122

Haastrup, K., 1987, Using Thinking Aloud and Retrospection to Uncover Learners' Lexical Inferencing Procedures (In Faerch, C. and Kasper, G. (Eds), 1987)

Haastrup, K., 1991, Lexical Inferencing Procedures or, Talking About Words, Tübingen:Gunter Narr Verlag

Haenggi, D. and Perfetti, C., 1994, Processing Components of College-Level Reading Comprehension, Discourse Processing 17:83-104

Halliday, M.A.K. and Hasan, R., 1976, Cohesion in English. London: Longman

Hamesse, J. and Zampolli, A. (Eds), 1985, Computers in Literary and Linguistic Computing, Paris: Champion-Slatkine

Harri-Augstein, S. and Thomas, L., 1984, Conversational Analysis of Reading: The Self-Organised Reader and the Text (In Alderson, J. and Urquhart, A., 1984)

Harris, D., 1985, Some Forerunners of Cloze Test Procedure, Modern Language Journal 69,4:367–376

Heaton, J.B., 1975, Writing English Language Tests, London: Longman

Henk, W. A., 1982, A Response to Shanahan, Kamil, and Tobin: The Case is Not Yet Clozed, Reading Research Quarterly 17:591-595

Henning, G., 1975, Measuring Foreign Language Reading Comprehension, Language Learning 25,1:109-114

Hendry, A. (Ed), 1982, Proceedings of the 18th Annual Conference of the United Kingdom Reading Association. London: Heinemann

Henry, G., 1979, The Relation Between Linguistic Factors Identified by a Principal Components Analysis of Written Style and Reading Comprehension as Measured by Cloze Tests, Journal of Reading Research 2,2:120-128

Hinofotis, F. and Snow, B., 1980, An Alternative Cloze Testing Procedure: Multiple-Choice Format (In Oller and Perkins (Eds), 1980)

Hopkins, E. and Grotjahn, R. (Eds), 1981, Studies in Language Teaching and Language Acquisition (Quantitative Linguistics 9) Bochum: Studienverlag Dr. N. Brockmeyer.

Hopkins, E. A., 1985, Redundancy, Entropy and the Comprehension Difficulty of Translation Texts: Groundwork for the Investigation of the Strategies of Adult German Learners of English (In Klein-Braley and Raatz (Eds) 1985)

Hornby, T. and Spann, M., 1981, ESL Reading Proficiency: Testing Strategies, On TESOL, 1981:25-33

Horn, R., Ingenkamp, K. and Jünger, R. (Eds), Tests Und Trends 1983, Jahrbuch Der Pädagogischen Diagnostik, Weinheim: Beltz

Hosenfeld, C., 1977, A Language-Teaching View of Second Language Instruction: The Learning Strategies of Second Language Learners with Reading-Grammar Tasks, Unpublished Phd Dissertation, Ohio State University, Columbus

Hughes, A. and Porter, D. (Eds), 1983, Current Developments in Language Testing, New York: Academic Press

Hughes, A., 1989, Testing for Language Teachers, Cambridge, CUP

Jacobs, E., 1987, Qualitative Research Traditions: A Review, Review of Educational Research 57,1:1-50

James, C., 1979, The Psychology of Cloze in Language Testing, Kongreßbericht der 10. Jahrestagung der Gesellschaft für Angewandte Linguistik E.V., Heidelberg: Julius Groos Verlag

Johnson, K. and Golombek, P. (Eds), 2002, Teachers' Narrative Inquiry as Professional Development , Cambridge; CUP

Johnston, P., 1983, Reading Comprehension Assessment: A Cognitive Basis, Newark, Del.: International Reading Association

Jonz, J., 1976, Improving the Basic Egg: The Multiple-Choice Cloze, Language Learning, 26,2:255-265

Jonz, J., 1987, Textual Cohesion and Second Language Comprehension, Language Learning 37:409-438

Jonz, J., 1989, Textual Sequence and Second Language Comprehension, Language Learning, 39,2:207-250

Jonz, J., 1990, Another Turn in the Conversation: What Does Cloze Measure?, TESOL Quarterly 24:61-83

Jonz, J., 1991, Cloze Item Types and Second Language Comprehension. Language Testing, 8,1:1-22.

Just, M. and Carpenter, P., 1977, Cognitive Processes in Comprehension, Hillsale N.J.: Erlbaum

Just, M. and Carpenter, P., 1980, A Theory of Reading: From Eye-Fixations To Comprehension, Psychological Review 87,4:329-354

Kamil, J., et al, 2000, A Handbook of Reading Research Vol. III, Newark NJ: National Reading Association

Kasper, G., 1984, Pragmatic Comprehension in Learner—Native Speaker Discourse, Language Learning 34:1-20

Katona, L. and Dörnyei, Z., 1993, The C-test: A Teacher-friendly Way to Test Language Proficiency. English Teaching Forum, 31,1: 34-35

Kelle, U. (Ed), 1995, Computer-Aided Qualitative Data Analysis, London: SAGE

Kennedy, A., 1978, Reading Sentences: Some Observations on the Control of Eye Movements (In Underwood, G. (Ed), 1978)

Kern, R. G., 1994, The Role of Mental Translation in Second Language Reading. Studies in Second Language Acquisition, 16:441-461.

Kesar, O., 1990. Identification and Analysis of Reading Motives in Completing a Rational Cloze, Unpublished MA Thesis, School of Education, Hebrew University, Jerusalem

Kintsch, W., 1982, Memory for Text (In Flammer, A. and Kintsch, W. (Eds), 1982)

Kintsch, W. and Van Dijk, T.A., 1978, Toward a Model of Text Comprehension and Production, Psychological Review 85,5:363-94

Kirby, J., 1988, Style, Strategy and Skill in Reading (In Schunker (Ed), 1988)

Klare, G.R. (1984), Readability (In Pearson, P. D. (Ed) 1984)

Klare, G.R., Sinaiko, H.W. and Stolurow, L.M., 1972, The Cloze Procedure: A Convenient Readability Test for Training Materials and Translations, International Review of Applied Psychology, 21,2: 77-106

Klein-Braley, C., 1980, The Assessment of Reading Comprehension Skills, Duisburg: University of Duisburg

Klein-Braley, C., 1981a, Empirical Investigations of Cloze Tests, Phd. Thesis, University of Duisburg

Klein-Braley, C., 1981b, The Determination of Text Difficulty, Paper Presented at the 12. Jahrestagung der Gesellschaft für Angewandte Linguistik, Mainz.

Klein-Braley, C., 1982, On The Suitability of Cloze Tests as Measures of Reading Comprehension (In Van Der Geest, A.J.M., Koster, C.J. and Matter, J.F. (Eds), 1982)

Klein-Braley, C., 1983, A Cloze is a Cloze is a Question (In Oller, J.W. Jr. (Ed), Issues in Language Testing Research. Rowley, Mass.: Newbury House.

Klein-Braley, C., 1984a, Advance Prediction of Difficulty with C-Tests (In Culhane, T., Klein-Braley, C. and Stevenson, D.K. (Eds), 1984)

Klein-Braley, C., 1984b, The Assessment of Reading Comprhension Skills in ESP Courses. AKS Rundbrief, 10: 20-37.

Klein-Braley, C., 1985a, A Cloze-Up on the C-Test: A Study in the Construct Validation of Authentic Tests. Language Testing, 2: 76-104.

Klein-Braley, C., 1985b, Advance Prediction of Test Difficulty. in: Klein-Braley and Raatz (Eds), 1985).

Klein-Braley, C., 1985c, C-Tests and Construct Validity (In Klein-Braley and Raatz (Eds), 1985)

Klein-Braley, C., 1985d, C-Tests as Placement Tests for German University Students of English (In Klein-Braley and Raatz (Eds), 1985)

Klein-Braley, C., 1985e, Reduced Redundancy as an Approach to Language Testing (In Klein-Braley and Raatz (Eds), 1985)

Klein-Braley, C., 1995, Appraisal of the C-Test in the Context of the Cloze Family (In Börner, W. & Vogel, K. (Eds) 1995) *Der Text im Fremdsprachenunterricht* (pp. 197-213). Bochum: AKS.

Klein-Braley, C. and Grotjahn, R., 1995, Der C-Test: Eine Eierlegende Wollmilchsau?, Paper Presented at the Congress of the Gesellschaft für Fremdsprachenforschung, Halle, 1995

Klein-Braley, C. and Raatz, U., (Eds), 1985, Fremdsprachen Und Hochschule 13/14: Thematischer Teil: C-Tests in Der Praxis

Klein-Braley, C. and Stevenson, D.K., (Eds), 1981, Practice and Problems in Language Testing, Frankfurt:Lang

Klenk, U. , Körner, K. and Thümmel, W. (Eds), 1989, Festschrift für G. Ineichen, Wiesbaden

Koda, K., 1988, Cognitive Process in Second Language Reading: Transfer of L1 Reading Skills and Strategies, Second Language Research 4,2:133-156

Koda, K., 1992, The Effects of Lower-level Processing Skills in FL Reading Performance: Implications for Instruction, The Modern Language Journal, 76,4:502-512

Koda, K., 1994, Second Language Reading Research: Problems and Possibilities, Applied Psycholinguistics 15,1:1-28

Krings, H., 1987, The Use of Introspective Data in Translation, (In Faerch, C. and Kasper, G., (Eds), 1987)

Krings, H., 1988, Blick in die `Black Box': Eine Fallstudie zum Übersetzungsprozeß bei Berufsübersetzern. in: Arntz, R. (Ed), 1988)

Kühlwein, W. and Raasch, A. (Eds), 1979, Sprache Und Verstehen Band 1 (Kongreßberichte Der 10 Jahrestagung Der GAL, Mainz 1979, Tubingen: Narr

Laberge, D. and Samuels, S. (Eds), Basic Processes in Reading: Perception And Comprehension, Hillsdale, NJ: Erlbaum)

Lado, R., 1986, Analysis of Native Speaker Performance on a Cloze Test, Language Testing 3,2:130-146

Lange. D. and Clausing, G., 1981, An Examination of Two Methods of Generating and Scoring Cloze Tests with Students of German on Three Levels, Modern Language Journal 65:254-261

Langer. J.A., 1984, Examining Background Knowledge and Text Comprehension, Reading Research Quarterly 19:468-481

Leow, R. and Morgan-Short, K., 2004, To Think Aloud or Not to Think Aloud: The Issue of Reactivity in SLA Research Methodology, SSLA 26:35–57

Little, D. and Singleton, D., 1992, The C-Test as an Elicitation Instrument in Second Language Research (In Grotjahn, R. (Ed), 1992)

Lorch, R., Lorch, E. and Mathews, P., 1981, On-Line Processing of the Topic Structure of a Text, Journal of Verbal Learning and Verbal Behaviour 19:350-369

Lunzer, E. and Gardner, K., 1979, (Eds), The Effective Use of Reading, London: Heinemann

Lutjeharms, M., 1988, Lesen in Der Fremdsprache: Versuch Einer Psycholinguistischen Deutung Am Beispiel Deutsch Als Fremdsprache, Bochum: AKS.

Lutjeharms, M. and Culhane, T. (Eds), 1982, Practice and Problems in Language Testing 3. Studiereeks Van Het Tijdschrift Van De Vrije Universiteit Brussel: Brussel

Macginitie, W., 1961, Contextual Constraint in Paragraphs, Journal of Psychology 51:121-130

Mackay, R., Barkman, N. and Jordan, R.R. (Eds), 1979, Reading in a Second Language. Rowley, Mass.: Newbury House

Mandl, H. (Ed), 1981, Zur Psychologie der Textverarbeitung: Ansaetze, Befunde, Probleme, München: Urban and Schwarzenberg

Mandl, H. and Ballstaedt, S.P., 1982, Effects of Elaboration on Recall of Texts, (In Flammer, A. and Kintsch, W. (Eds), 1982)

Mandl, H., Stein, N. and Trabasso, T., 1984, Learning and Comprehension of Text, Hillsdale, NJ: Erlbaum

Mangubhai, F., 1990, Towards a Taxonomy of Strategies Used by ESL Readers of Varying Proficencies While Doing Cloze Exercises, Australian Journal of Reading 13,2:128-139

Mannes, S., 1994, Strategic Processing of Text, Journal of Educational Psychology 86,4:577-588

Markham, P., 1987, Rational Deletion Cloze Processing Strategies: ESL and Native English, System 15,3:303-311

Markham, P., 1988, The Cloze Procedure and Intersentential Comprehension in College-Level German, IRAL 26,1:44-51

Matsumoto, K. , 1993, Verbal-Report Data and Intropective Methods in Second Language Research: State of the Art. RELC Journal 24:32-60

Maynard, S., 1997, Japanese Communication, Honolulu: University of Hawai'i Press

McKenna, M., 1978, Cumulative Interference of Incorrect Cloze Responses, Reading Improvement 15,3:172-174

McLeod, B. and McLaughlin, B., 1986, Restructuring or Automaticity?: Reading in a Second Language, Language Learning 36:109-123

McNamara, T., 2000, Language Testing, Oxford: OUP

Meredith, K. and Vaughan, J., 1978, Stability of Cloze Scores Across Varying Deletion Patterns (In: Pearson, P.D. and Hansen, J. (Eds), 1978)

Miles, M. and Haberman, A., 1984, Qualitative Data Analysis: A Sourcebook of New Methods Beverley Hills, CA: Sage

Miller, C., 1956, The Magical Number Seven, Plus Or Minus Two: Some Limits on Our Capacity for Processing Information, The Psychological Review 63,2:81–97

Miller. G. and Selfridge, J., 1950, Verbal Context and the Recall of Useful Material, American Journal of Psychology 63:176–185

Mogge, B., 1979, Probleme der Verständlichkeit in Gebrauchstexten (In Kühlwein, W. and Raasch, A. (Eds), 1979)

Naiman,N., et al, 1978, The Good Language Learner, Toronto, Ont.: Institute For Studies in Education

Nattinger, J.R. and Decarrico, J., 1992, Lexical Phrases and Language Teaching, Oxford:OUP

Neumann, G., 1995, Laut Denken Und Still Schreiben: Zur Triangulierung Von Prozeß-Und Produktdaten in Der L2 Schreibprozeßforschung, Zeitschrift für Fremdsprachenforschung 6,1:95-107

Neville, M.H. and Pugh, A.K., 1976, Context in Reading and Listening: Variations in Approach to Cloze Tasks. Reading Research Quarterly, 12,1:13-31.

Nevo, N., 1989, Test-Taking Strategies on a Multiple-Choice Test of Reading Comprehension, Language Testing 6,2:199-215

Nisbett, R.E. and Wilson, T.D., 1977, Telling More Than We Can Know: Verbal Reports on Mental Processes, Psychological Review 84: 231-59

Nord, C., 1989, Textanalyse Und Übersetzen. Heidelberg, Julius Groos.

Nuttall, C.. 1982, Teaching Reading Skills in a Foreign Language. London: Heinemann.

Ohnmacht, F.W., Weaver, W.S. and Kohler, E.T., 1970, Cloze and Closure: A Factorial Study, Journal of Psychology, 14,:205-217.

Oller, J.W.Jr., 1973, Pragmatic Language Testing, Language Sciences, 28:7-12

Oller, J.W.Jr., 1974, Expectancy for Successive Elements: Key Ingredient to Language Use, Foreign Language Annals 7:443-452

Oller, J.W. Jr., 1975a, Cloze, Discourse and Approximations to English (In Burt, M.K. and Dulay, H.C. (Eds), 1975)

Oller, J.W. Jr., 1975b, Research with Cloze Procedure Relevant to the Investigation of the Proficiency of Nonnative Speakers of English, Washington, D.C.: Center For Applied Linguistics

Oller, J.W. Jr., 1976a, Language Testing (In Wardhaugh, R. and Brown, H. (Eds), 1976)

Oller, J.W. Jr., 1976b, Evidence for a General Language Proficiency Factor: An Expectancy Grammar, Die Neueren Sprachen, 75:165-174

Oller, J.W. Jr., 1979, Language Tests at School: A Pragmatic Approach, London: Longman

Oller, J.W.Jr., 1981, Language as Intelligence? Language Learning, 31,2:465-492.

Oller, J.W. Jr., 1983, Issues in Language Testing Research, Rowley, MA: Newbury House

Oller, J.W.Jr., Bowen, D., Dien, T.T. and Mason, V.W., 1972, Cloze Tests in English, Thai and Vietnamese: Native and Non-Native Performance, Language Learning, 22,1:1-15

Oller, J.W.Jr. and Conrad, C., 1971, The Cloze Procedure and ESL Proficiency. Language Learning, 21:183-195

Oller, J.W. Jr., and Hinofotis, F., 1980, Two Mutually Exclusive Hypotheses About Second Language Ability: Indivisible or Partially Divisible Competence (In Oller and Perkins (Eds), 1980)

Oller, J.W.Jr. and Jonz, J., 1994, Cloze and Coherence, Lewisburg, PA: Bucknell University Press

Oller, J.W.Jr. and Nevin, I., 1971, A Cloze Test of English Prepositions, TESOL Quarterly, 5:315-326

Oller, J.W.Jr. and Perkins, K. (Eds), 1978, Language in Education: Testing the Tests, Rowley, Mass.: Newbury House

Oller, J.W. Jr. and Perkins, K. (Eds), 1980, Research in Language Testing, Rowley, MA: Newbury House

Oller, J.W. Jr. and Streiff, V., 1975, Dictation: A Test of Grammar-Based Expectancies, English Language Teaching, 30:25-35

Olshavsky, J., 1977, Reading as Problem Solving: An Investigation of Strategies, Reading Research Quarterly, 12:654-674

O'Malley, J. et al, 1985, Learning Strategy Applications with Students of English as a Second Language, TESOL Quarterly 19,3:557-584

O'Malley, J. and Chamot, A., 1990, Learning Strategies in Second Language Acquisition, Cambridge: CUP

Osgood, C., 1959, The Representational Model (In De Sola Pool, I. (Ed), 1959)

Osgood, C. and Sebeok, T., (Eds), 1965, Psycholinguistics: A Survey of Theory and Research Problems, Bloomington, IN: University of Indiana Press

Oxford, R. L., 1990, Language Learning Strategies; What Every Teacher Should Know, NY: Newbury House/Harper Collins

Pearson, P. and Hansen, J. (Eds), 1978, Reading: Disciplined Inquiry in Process and Practice (27th Yearbook of the National Reading Conference), Clemson, S.C.

Pearson, P. D. (Ed), 1984, Handbook of Reading Research, New York and London: Longman

Perfetti, C., Goldman, S. and Hogaboam, T., 1979, Reading Skill and the Identification of Discourse Context, Memory and Cognition 7:273-282

Perfetti, C.A. and Goldman, S.R., 1976, Discourse Memory and Reading Comprehension Skill, Journal of Verbal Learning and Verbal Behavior, 15:33-42.

Perkins, K., 1992, The Effect of Passage and Topical Structure Types on ESL Reading Comprehension Difficulty. Language Testing 9:163-172

Perkins, K. and James, B., 1985, Measuring Passage Contribution to EFL Reading Comprehension, TESOL Quarterly 19:137-153

Perren, G.E. and Trim, J.L.M. (Eds), 1971, Applications of Linguistics, Cambridge: Cambridge University Press.

Petersen, C., 1992, Identifying Referents and Linking Sentences Cohesively in Narration, Discourse Processes 16:507-524

Porter, D., 1976, Modified Cloze Procedure; A More Valid Reading Comprehension Test, English Language Teaching 30:151-155

Porter, D., 1978, Cloze Procedure and Equivalence, Language Learning 28:333-341

Porter, D., 1983, The Effect of Quantity of Context on the Ability to Make Linguistic Predictions (In Hughes, A, and Porter, D. (Eds) 1983)

Poulisse, N., Bongaerts, T. and Kellerman, E., 1987, The Use of Retrospective Verbal Reports in the Analysis of Compensatory Strategies (In Faerch, C. and Kasper, G., (Eds), 1987)

Pressley, M. and Afflerbach, P.,1995, Verbal Protocols of Reading: The Nature of Constructively Responsive Reading, Hillsdale, N.J.: Erlbaum

Pugh, A.K., 1981a, Construction and Reconstruction of Text (In: Chapman, L.J. (Ed), 1981)

Pugh, A.K., 1981b, Text Analysis, Comprehension and Reading Behaviour (In Hopkins, E. and Grotjahn, R. (Eds), 1981)

Pugh, A.K., 1982, Practical Applications of Analyses and Descriptions of Texts (in Hendry, A. (Ed), 1982)

Pugh, A.K., 1983, Development of Fluency and Strategy in Native and Foreign Language Reading: Some Comparisons and Contrasts, Paper Presented To the Third European Conference On Reading, Vienna

Pugh, A.K., 1985, Computer Analysis of Cloze Responses: A Comprehensive System and Some Findings From Its Application (In Hamesse, J. and Zampolli, A. (Eds), 1985)

Pugh, A.K. and Ulijn, J.M., 1981, Some Approaches to Studying Realistic Reading Tasks, Paper Presented to the 1981 International Symposium On Text Processing, Fribourg, Switzerland

Purpura, J., 1997, An Analysis of the Relationships Between Test Takers' Cognitive and Metacognitive Strategy Use and Second Language Test Performance, Language Learning 42,2:289–325

Raatz, U., 1982, Language Theory and Factor Analysis (In Lutjeharms, M. and Culhane, T. (Eds), 1982)

Raatz, U., 1984, The Factorial Validity of C-Tests (In Culhane, T., Klein-Braley, C. and Stevenson, D.K. (Eds), 1984)

Raatz, U, 1985, Tests of Reduced Redundancy – The C-Test, A Practical Example (In Klein-Braley and Raatz (Eds), 1985)

Raatz,U. and Klein-Braley, C., 1981, The C-Test – A Modification of the Cloze Procedure (In Culhane, T., Klein-Braley, C. and Stevenson, D.K. (Eds), 1984)

Raatz, U. and Klein-Braley, C., 1983, Ein Neuer Ansatz zur Messung der Sprachleistung, Der C-Test: Theorie Und Praxis (In Horn, R., Ingenkamp, K. and Jünger, R. (Eds), 1983)

Raatz, U. and Klein-Braley, C., 1985, How to Develop a C-Test (In Klein-Braley and Raatz (Eds), 1985)

Ramanauskas, S., 1972, The Responsiveness of Cloze Readability Measures to Linguistic Variables Operating Over Segments of Text Longer Than a Sentence, Reading Research Quarterly 8:72-91

Rand, E., 1978, The Effects of Test Length and Scoring Method on the Precision of Cloze Test Scores, Workpapers in Teaching English as a Second Language 12:62–71, Los Angeles: University of California, LA

Rankin, E.F.Jr. and Culhane, J.W., 1959, Comparable Cloze and Multiple-Choice Comprehension Test Scores, Journal of Reading, 13:193-198.

Rayner, K. and Morris, R., 1990, Do Eye Movements Reflect Higher Order Processes in Reading? (In Groner, R., d'Ydewalle, G. and Parham, R. (Eds), 1990)

Rayner, K. and Pollatsek, A., 1989, The Psychology of Reading, Englewood Cliffs, NJ:Prentice-Hall

Read, J., 2000, Assessing Vocabulary, Cambridge:CUP

Ridley, J., 1997, Reflections and Strategies in Foreign Language Learning, Frankfurt am Main: Peter Lang

Robinson. P., 2001. Task Complexity, Task Difficulty, and Task Production: Explaining Interactions in a Componential Framework, Applied Linguistics 22,1:27-57

Rye, J. 1979, A Closer Look At Cloze, English in Education, 13,3:44-54.

Rye, J., 1984, Cloze and Intersentential Constraint, Journal of Reading Research 7,2:113-122

Sasaki, M., 1993, Relationships Among 2nd-Language Proficiency, Foreign-Language Aptitude, and Intelligence: A Structural Equation Modeling Approach, Language Learning 43:313-344

Schunker, R. (Ed), 1988, Learning Strategies and Learning Styles, N.Y.: Plenum

Sciarone, A. and Schoorl, J., 1989, The Cloze Test: Or Why Small Isn't Always Beautiful, Language Learning 39:415-438

Seidel, J. and Kelle, U., 1995, Different Functions of Coding in the Analysis of Textual Data (In Kelle, U. (Ed), 1995)

Shanahan, T., Kamil, M. and Tobin, A., 1982, Cloze as a Source of Intersentential Comprehension, Reading Research Quarterly 17:229-254

Shannon, C. E. and Weaver, W., 1949, The Mathematical Theory of Communication, Urbana: University of Illinois Press

Sharwood-Smith, M., 1979, Strategies, Language Transfer and the Simulation of the Second Language Learner's Mental Operations. Language Learning, 29,2:345-361

Sierra, M. (Ed), 1994, Languages in the European Community, Amsterdam: Rodopi

Sim, D. and Bensoussan, M., 1979, Control of Contextualized Function and Content Words as It Affects EFL Reading Comprehension Test Scores (In Mackay, R., Barkman, N. and Jordan, R.R. (Eds), 1979)

Simons & Lodewijks 1988

Simpson, B., 1912, Correlations of Mental Abilities, Columbia University Contributions to Education (Reprint, New York:AMS Pres 1972)

Skehan, P., 1989, Individual Differences in Second Language Learning, London: Arnold

Smagorinsky, P., 1998, Thinking and Speech and Protocol Analysis, Mind, Culture and Activity 5:157–177

Smith, F., 1978, Understanding Reading (3rd ed.) New York: Holt, Rinehart and Winston

Smith, V., 1994, Thinking in a Foreign Language: An Investigation Into Essay Writing and Translation By L2 Learners, Tubingen: Gunter Narr

Smith-Burke, M., Gingrich, P.S. and Eagleye, D., 1978, Differential Effect of Prior Context, Style, and Deletion Pattern On Cloze Comprehension (In Pearson and Hansen 1978:133–137)

Souza, M., 1994, Schemata Theory and Lexical Inference in Reading English Text, Mimesis 15,1:163-177

Spolsky, B., Bengt, S.M., Sako, E.W. and Aterburn, C., 1968, Preliminary Studies in the Development of Techniques for Testing Overall Second Language Proficiency. (In Upshur, J.A. and Fata, J. (Eds), 1968)

Spolsky, B., 1971, Reduced Redundancy as a Language Testing Tool (In Perren, G.E. and Trim, J.L.M. (Eds), 1971)

Spolsky, B., 1981, Some Ethical Questions About Language Testing (In Klein-Braley, C. and Stevenson, D.K., (Eds), 1981)

Spolsky, B., 1989, Communicative Competence, Language Proficiency, and Beyond, Applied Linguistics, 10,2:138-156.

Stevenson, D.K., 1981, `All of the Above': On Problems in the Testing of Foreign Language Reading. System, 9,3:267-273.

Stevenson, D.K., 1979, Face Validity and Loss of Faith, Paper Presented to the Conference of the Gesellschaft Für Angewandte Linguistik, Mainz.

Stevenson, D.K., 1979, Face Validity and Loss of Faith: Some Effects of Recent Cloze Research on Traditional Views of Language Proficiency, Kongreßbericht Der 10.Jahretagung D. Gesellschaft für Angewandte Liguistik FAL E. V., Heidelberg: Julius Groos Verlag

Taft, M., 1991, Reading and the Mental Lexicon, Hillsdale, NJ: Erlbaum

Tanenhaus, M.K. and Lucas, M.M., 1987, Context Effects in Lexical Processing. Cognition, 25:213-234.

Taylor, W.L, 1953, Cloze Procedure: A New Tool for Measuring Readability. Journalism Quarterly 30:415-433.

Taylor, W.L., 1956, Recent Developments in the Use of `Cloze Procedure', Journalism Quarterly, 33:42-48

Taylor, W.L., 1957a, `Cloze' Readability Scores as Indices of Individual Differences in Comprehension and Aptitude. Journal of Applied Psychology, 41: 19-26.

Tergan, S.O., 1981, Ist Textverstaendlichkeit Gleich Textverstaendlichkeit? (In Mandl, H. (Ed), 1981)

Trabasso, T. and Suh, S., 1993, Understanding Text: Achieving Explanatory Coherence Through Online Inferences and Mental Operations in Working Memory, Discourse Processes 16,1-2:3-34

Trabue, M.R., 1916, Completion-Test Language Scales, New York: Teachers' College, Columbia University

Trollope, J. 1995, A Comparison of Reading Strategies in a Computerized and in a Non-Computerized Reading Situation, Unpublished Master's Thesis, Univeristy of Illinois, Retrieved from the Web Oct. 14th, 2000 (http://www.iei.uiuc.edu/resources)

Tuinman, J.J. and Gray, G., 1972, The Effect of Reducing the Redundancy of Written Messages by Deletion of Function Words, Journal of Psychology, 82:299-301.

Tuinman, J.J., 1974, Determining the Passage Dependency of Comprehension Questions in 5 Major Tests, Reading Research Quarterly, 9,2:206-223.

Tuinman, J.J., Blanton, W.E. and Gray, F., 1975, A Note on Cloze as a Measure of Comprehension, Journal of Psychology, 90:159-162.

Underwood, G. 1977, Attention, Awareness and Hemispheric Differences in Word Recognition, Neuropsychologia 15:61-67

Underwood, G., 1978 (Ed), Strategies of Information Processing, London: Academic Press

Upshur, J.A. and Fata, J. (Eds), 1968, Problems in Foreign Language Testing, Language Learning Special Issue No. 3

Upton, J., 1997 First and Second Language Use in Reading Strategies of Japanese ESL Students, TESL-EJ 3,1, Retrieved from the Web July 8th, 2002 (http://www-writing.berkeley.edu/TESL-EJ/ej09/a3.html)

Van De Velde, R., 1992, Text and Thinking: On the Roles of Thinking in Text Interpretation, Berlin: De Gruyter

Van Den Brandt, C. and Van Esch, K., 1994, A New Test of Written Comprehension in the Teaching of Spanish as a Foreign Language, (In Sierra, M. (Ed), 1994)

Van Der Geest, A.J.M., Koster, C.J. and Matter, J.F. (Eds), 1982, Toegepaste Taalwetenschap in Artikelen 13, Amsterdam: VU Boekhandel.

Van Dijk, T.A. and Kintsch, W., 1983, Strategies of Discourse Comprehension. New York: Academic Press.

Van Parreren, C. and Shouter-Van Parreren, M., 1981, Contextual Guessing: A Trainable Reader Strategy, System 9,3:235-241

Von Faber, Kreifelts et al, 1978, Kongreßberichte Der 9. Jahrestag Der GAL Band III Heidelberg: Julius Groos

Vygotsky, L.S., 1978, Mind in Society: The Development of Higher Psychological Processes, Cambridge, MA: Harvard University Press

Vygotsky, L.S., 1986, Thought and Language, Cambridge, MA: MIT Press

Waern, Y., 1982,, How Do You Fill This xxx?: On Some Interpretation Processes (In Flammer, A. and Kintsch, W. (Eds), 1982)

Waern, Y., 1988, Thoughts on Text in Context: Applying the Think-Aloud Method to Text Processing, Text 8:327–350

Wales, R. and Walker, E. (Eds), 1976, New Approaches to Language Mechanisms, Amsterdam: North Holland

Wardhaugh, R. and Brown, H. (Eds), 1976, A Survey of Applied Linguistics, Ann Arbor, MI; University of Michigan Press

Wenden, A., 1986, What Do Learners Know About Their Language Learning?, Applied Linguistics 7:186-201

Werner, H., 1954, Change of Meaning: A Study of Semantic Processes Through the Experimental Method, Journal of General Psychology 50:181-208

Whalen, K. and Menard, N., 1995, L1 and L2 Writer's Strategic and Linguistic Knowledge: A Model of Multiple-Level Discourse Processing, Language Learning 45,3:381-418

White, P., 1980, Limitations on Verbal Reports of Internal Events: A Refutation of Nisbett and Wilson and of Benn, Psychological Review 87,1:105–112

Widdowson, H.G., 1989, Knowledge of Language and Ability for Use, Applied Linguistics, 10,1:128-137.

Wolfe-Quintero, K., S. Inagaki and H.-Y. Kim, 1998, Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity, Manoa: University of Hawaii

Wolff, D., 1989, Identification of Text-Type as a Strategic Device in L2 Comprehension (In Dechert and Raupach (Eds), 1989)

Wolff, D., 1993, A Comparison of Assessment Tasks Used to Measure FL Reading Comprehension, MLJ 77,4:473-489

Yi'an, W., 1998, What Do Tests of Listening Comprehension Test? A Retrospection Study of EFL Test-Takers Performing a Multiple-Choice Task, RCEAL Working Papers 2:75-89

Zettersten, A., 1977, Papers on English Language Testing in Scandanavia, Copenhagen: University of Copenhagen

Zimmermann, R., 2000, L2 Writing Subprocesses: A Model of Formulating and Empirical Findings, Learning and Instruction 10:73-99, Pergamon

Zwaan, R., 1994, Effect of Genre Expectations on Text Comprehension, Journal of Experimental Psychology (Learning, Memory and Cognition) 20,4:920-933

Appendix 1: Olympics Text & Cloze Passage

APPENDIX 1

(1) Instructions To TA & AC Informants (TL version);
(2) OLYMPICS text (abbreviated);
(3) OLYMPICS Cloze Task Passage (32-deletion)

INSTRUCTIONS TO INFORMANTS

Thank you for coming today! Under this page is a 'cloze' passage about the Olympic Games. Every 7th word has been replaced by a blank space. In this space you have to write *one* English word which makes sense there.

If you can think of more than one possible word which could go in the space, please choose the one you think fits best. Each space requires one, but *only one*, word.

When you feel ready, begin trying to fill the blanks. As you do so, please

[TA condition] try to explain your thinking processes onto your tape, as you did in the practice task. What information, from the text or elsewhere, are you using to help you choose the best words? What actions are you taking as you work on the passage?

[AC condition] try to show your thinking process by noting down suitable codes from the set you were introduced to in the orientation and the practice task. What information, from the text or elsewhere, are you using to help you choose the best words? What actions are you taking as you work on the passage?

If you have any questions about what to do, please ask now!

You do not have to fill the blanks in the order in which they appear on the page. You may take as much time of this class / session time as you need to complete the test, or as much of it as you feel like doing. Although there is no time limit, however, please try to pretend the task is a real test.

Remember that no-one else but me will know your score, but you may come and ask me individually about your *own* score, and about why a word is right or wrong here.

I will take notes as you work through the test, but I will not interrupt you except perhaps to ask what you are thinking or working on at that moment.

Please do not talk to others in the room, and please do not ask me questions during the task. You may leave the room if necessary without asking permission, but please come back as soon as possible.

'OLYMPICS' PASSAGE (ABBREVIATED)

In ancient Greece athletic festivals were very important and had strong religious associations. The Olympian athletic festival, held every four years in honour of Zeus, eventually lost it's local character, became first a national event, and then, after the rules against foreign competitors had been waived, international. No one knows exactly how far back the Olympic Games go, but some official records date from 776 B.C. The Games took place in August on the plain by Mount Olympus. Many thousands of spectators gathered from all parts of Greece, but no married woman was admitted even as a spectator. Slaves, women and dishonoured persons were not allowed to compete. The exact sequence of event is uncertain, but events included boys' gymnastics, horse-racing, field events such as discus and javelin throwing, and the very important foot races.

There was also boxing and wrestling and special tests of varied ability such as the pentathlon, the winner of which excelled in running, jumping , discus and javelin throwing and wrestling. The evening of the third day was devoted to sacrificial offerings to the heroes of the day, and the fourth day, that of the full moon, was set aside as a holy day. On the sixth and last day, all the victors were crowned with holy garlands of wild olive from a sacred wood. So great was the honour that the winner of the foot race gave his name to the year of his victory!

'OLYMPICS' CLOZE PASSAGE

(32 Deletion Abbreviated Version)

In ancient Greece athletic festivals were very important and had strong religious associations. The Olympian athletic festival, held every (1)_____ years in honour of Zeus, eventually (2)_____ its local character, became first a (3)_____ event, and then, after the rules (4)_____ foreign competitors had been waived, international. (5)_____ one knows exactly how far back (6)_____ Olympic Games go, but some official (7)_____ date from 776 B.C.

The Games (8)_____ place in August on the plain (9)_____ Mount Olympus. Many thousands of spectators (10)_____ from all parts of Greece, but (11)_____ married woman was admitted even as (12)_____ spectator. Slaves, women and dishonoured persons (13)_____ not allowed to compete. The exact (14)_____ of events is uncertain, but events (15)_____ boys' gymnastics, horse-racing, field events (16)_____ as discus and javelin throwing, and (17) _____ very important foot races.

There was (18)_____ boxing and wrestling and special tests (19)_____ varied ability such as the pentathlon, (20)_____ winner of which excelled in running, (21)_____ , discus and javelin throwing and wrestling. (22)_____ evening of the third day was (23)_____ to sacrificial offerings to the heroes (24)_____ the day, and the fourth day, (25)_____ of the full moon, was set (26)_____ as a holy day. On the (27)_____ and last day, all the victors (28)_____ crowned with holy garlands of wild (29)_____ from a sacred wood. So great (30)_____ the honour that the winner of (31)_____ foot race gave his name to (32)_____ year of his victory!

APPENDIX 2 Codes used in analysing TA protocols and AC manuscripts (TL version)

Codes denoting behaviours that precede or 'interrupt' recovery:

Rprt(a)

The informant reads part of passage (aloud) before beginning to recover all or some cloze deletions.

Rall(a)

The informant reads the entire passage (aloud) before beginning to recover cloze deletions.

Codes denoting use of local item context information:

Gra

The informant refers to or applies knowledge of a TL syntactic ('grammar') structure of which the missing items may be part (article--noun sequence, auxiliary verb--main verb, etc.) or which otherwise conditions or explains the choice of filler.

Phr

The informant refers to or applies knowledge of a TL phrase of which the missing item may be part.

Col

The informant refers to or applies knowledge of a TL collocation to which the missing item may belong.

Idi

The informant refers to or applies knowledge of a TL idiomatic expression of which the missing item may be part.

Codes denoting use of non-local cues

SS (SSb, SSf)

The informant refers to information contained within the same sentence

(SSb preceding the deletion; SSf following the deletion.)

SPb

The informant refers to passage information from within the same paragraph, but in a sentence preceding that containing the deletion.

SPf

The informant refers to passage information from within the same paragraph, but in a sentence following that containing the deletion.

EP

The informant refers to passage information from an earlier paragraph to that containing the deletion.

LP

The informant refers to passage information from a later paragraph to that containing the deletion.

KOW

The informant refers to her prior/extratextual knowledge of the world.

LAN

The informant refers to her knowledge of, or poses a question about, some aspect of the structure of the TL, her L1, or an L3

L1Par

The informant refers to a parallel or (near-)equivalent phrase or structure in her L1 (**L3Par** if the phrase or structure is in a third language)

Codes indicating partial success:

WC

The informant cites a word class (noun, verb, etc.) to which the target item must *or cannot* belong.

L1Eq

The informant inserts an appropriate L1 equivalent for a TL filler.

Und

The informant claims or in some way demonstrates understanding of the meaning of the missing item or its immediate context, but cannot produce a suitable filler.

SWA

The informant indicates that she is seeking a *single-word* filler for the deletion.

Codes denoting 'routes' to recovery or comprehension:

???

The informant either 'passively' gives no indication of how she recovered a fillerword, or states that she can give no indication.

Log

The informant makes a logical deduction or inference apparently based on passage information.

LNK

The informants appears to process 2 or more deletions together, or to link them within a single span of passage.

SKP

The informants reads through the passage at speed, apparently filling only those blanks whose fillers can be recovered quickly or easily.

GS

The informant indicates that she is guessing at appropriate filler word.

Codes indicating processing in L1:

Tr

The informant translates or glosses passage information within the same sentence.

Tr+

The informant translates or glosses passage information beyond the sentence.

UNDERLINED PASSAGE TEXT (specific to AC)

The informant translates this span of the passage into her L1.

Codes denoting evaluation or selection among fillers:

CBAs

The informant indicates that she is attempting to choose between two or more candidate fillers.

JSR

The informant indicates that a filler 'just seems right', i.e. she does not mention sound or appearance.

Sou

The informant explicitly refers to 'sound' of a filler in evaluating it.

LK

The informant explicitly refers to the 'look' of a filler in evaluating it.

Codes indicating task difficulty:

RF

The informant inserts a temporary 'running filler' such as a TL or L1 equivalent of 'something', a sound, etc.

UPI

The informant indicates unfamiliarity with an extant passage word, phrase, or structure.

+Diff

The informant indicates considerable difficulty in thinking of a suitable filler, or in comprehending an extant passage span.

Nec?

The informant questions whether a filler is actually necessary.

LFN

The informant announces that she will leave the target item for now.

GUP

The informant announces that she is giving up on the target item.

CL

The informant indicates that her confidence in her chosen filler is low

Codes indicating perception of production or comprehension error:

Cha

The informant indicates that she is changing a previous answer.

Ah!

The informant indicates that her understanding of (part of) the passage has altered.

made ah [INAUDIBLE] I GET IT.. made the *hard* training not the first I DON'T LIKE THAT ANY MORE.. GL1 Dm

Codes specific to pair-condition reporting:

Ptnr

The informant indicates that her partner has produced the answer.

CP

The informant consults with her partner or seeks her opinion about a filler or meaning.

Did

The informant appears to be teaching her partner something (she assumes to be) unknown to her

Diff

The informant indicates that she and partner have different answers at that point.

Defr

One informant defers to her partner's choice of filler, understanding of the passage, etc.

Solo

The informant fills a deletion without consulting her partner, or overtly seeking her agreement

Other Codes:

Rev

The informant indicates that she will review all or part of her processing of the task to date, or return to a given item.

Par

The informant paraphrases part of the passage in the TL or summarises her understanding of it.

Codes related to informant-researcher interactions:

BRK

Informant indicates that she will interrupt the session

Codes, etc. specific to AC reporting:1 2 3 4 5

The informant indicates the level of difficulty she had in finding an answer by choosing a scale value.

1 2 3 4 5

The informant indicates the level of confidence she has in her answer by choosing a scale value.

Graphic markups on protocols/manuscripts:

(Score through)

The informant has abandoned the deleted filler (revisions may be written in above)

(arrowed) line, boxing, circling, etc.

The informant identifies an association between spans of passage, or between fillers

underlined passage span

In AC indicates passage content translated into L1

L1 gloss

L1 meaning of extant passage item inserted (to date only noted on JL1 informants' task sheets and

Appendix 3: Sample AC manuscript

The Olympic Games

In ancient Greece athletic festivals were very important and had strong religious associations. The Olympian athletic festival, held every (1) 4 years in honour of Zeus, eventually (2) lost its local character, became first a (3) national event, and then, after the rules (4) forbidding foreign competitors had been waived, international (5) No one knows exactly how far back (6) the Olympic Games go, but some official (7) ~~claims~~ assigned date from 776 BC.

The Games (8) took place in August on the plain (9) near Mount Olympus. Many thousands of spectators (10) came from all parts of Greece, but (11) no married woman was admitted even as (12) a spectator. Slaves, women and dishonoured persons (13) were not allowed to compete. The exact (14) number of events is uncertain, but events (15) included boys' gymnastics, ~~horse-racing,~~ field events (16) such as discus and javelin throwing, and (17) other very important foot races.

There was (18) also boxing and wrestling and special tests (19) with varied ability such as the pentathlon, (20) the winner of which excelled in running, (21) jumping, discus and javelin throwing and wrestling. (22) The evening of the third day was (23) dedicated to sacrificial offerings to the heroes (24) of the day, and the fourth day, (25) _____ of the full moon, was set (26) side as a holy day. On the (27) first and last day, all the victors (28) were crowned with holy garlands of wild (29) _____ from a sacred wood. So great (30) was the honour that the winner of (31) the foot race gave his name to (32) the year of his victory!

1 (2) 3 4 5 — 1 — KOW

1 2 (3) 4 5 — 2 — WC

1 2 (3) 4 5 — 3 — SSf

1 2 3 (4) 5 — 4 — Phr, Tr

(1) 2 3 4 5 — 5 — Phr

1 (2) 3 4 5 — 6 — Gra

1 2 3 4 (5) — 7 — WC

(1) 2 3 4 5 — 8 — Phr

1 (2) 3 4 5 — 9 — Gra

(1) 2 3 4 5 — 10 — Phr

(1) 2 3 4 5 — 11 — KOW

(1) 2 3 4 5 — 12 — gra

(1) 2 3 4 5 — 13 — gra

1 2 (3) 4 5 — 14 — WC, phr

1 2 3 (4) 5 — 15 — Tr

(1) 2 3 4 5 — 16 — Phr

1 2 (3) 4 5 — 17 — Tr

(1) 2 3 4 5 — 18 — SSf, Tr

1 2 (3) 4 5 — 19 — Tr

1 (2) 3 4 5 — 20 — gra

1 2 (3) 4 5 — 21 — KOW 規律-

1 (2) 3 4 5 — 22 — gra

1 (2) 3 4 5 — 23 — Tr

(1) 2 3 4 5 — 24 — gra

1 2 3 4 5 — 25 — WC

(1) 2 3 4 5 — 26 — Phr

1 2 3 4 (5) — 27 — 形容詞 WC (さらに)規律-

(1) 2 3 4 5 — 28 — gra

1 2 3 4 5 — 29 — LIEg, Tr 月桂樹

(1) 2 3 4 5 — 30 — gra

1 2 (3) 4 5 — 31 — gr. Tr.

1 (2) 3 4 5 — 32 — gra. Tr

Appendix 4: Detailed comparisons of the processing of 10 passage items under think aloud, NCR and AC conditions

DELETION (11) [NO]

GL1 think aloud Tom

"...um perhaps *a* *married woman* ..that is a married woman or umm.. no rubbish! it's *married woman.. as a ..was admitted as..* [fills (12)].. *no married woman was admitted ...*"

Cha, Gra

GL1 think aloud Stella

"...*but even as* not [...] *no married woman* not even as *no married...*"

??? (--)

GL1 AC Andy

Log; SSb (NO)

GL1 AC Katrin

Gra (NO)

JL1 think aloud Kiyoko

"...eh *married woman was admitted even as was no* ummm ... *women* [...] only this.. eh?.. strange no? *but all parts of Greece.. but..only married woman..* mm? *only women..* no.. ah.. *no married woman.. was admitted...*"

+Diff/UPI; Ah!

JL1 think aloud Hideko

"...*but mmm people married woman was as admitted people greece greek mmm.. married women was admitted ...*"

"...*but men many woman woman* eh?.. *was* so it's.. *man* maybe? *man.. married woman was admitted* [...] what? to look? old.. old.. to do that.. nope.. well.. Greek.. *as..* [...] *Greece.. Greek married woman was admitted...*"

Sou (--)

JL1 NCR Jun

"*no married woman was admitted.. even as a spectator..* that's just grammar "

Gra (NO)

JL1 NCR Miho

"I remember learning this.. it means that women were not allowed even to watch.."

Gra (NO)

JL1 AC Ippei

Gra (NO)

JL1 AC Yuna

SSb; Gra (NO)

Comments on deletion (11)

While both GL1 think aloud informants filled the deletion appropriately, neither protocol gives any real insight into how they arrived at the correct filler. In PTIs, Tom claimed to have "more or less translated" in his head, while Stella claimed to have used her knowledge of grammar, as well as the [rhetorical?] relationship between 'but' and its following clause. She made no comment on her mention of 'not even' in this extract.

AC informant Andy coded for use of logical inference, and for unspecified information earlier in the sentence. (This is presumably the 'but' noted by Stella.)

AC informant Katrin coded only for grammatical knowledge, but had also circled 'even as' later in the sentence and linked it with a line to deletion (11). Even though the coding SSF does not appear, it seems plausible that the relation 'NEG+even' was in some way relevant.

Kiyoko's protocol offers no unambiguous insight into how she filled the blank, although she appears to have experienced some difficulty in making sense of the context, or found an unfamiliar structure. She then seems to have suddenly reinterpreted the context and arrived at the correct filler. Post-task, Kiyoko could offer no insight into her processing of this item, even as to whether she had filled the blank on the basis of grammatical knowledge. Hideko failed to fill the deletion correctly, and appeared to gave evaluated candidate fillers by sound. Post-task, she could only add that 'Greek' seemed an appropriate filler, and reacted with apparent surprise when shown the original word.

Jun and Miho both made explicit reference to grammar rules, which Jun, too, almost certainly acquired in school.

Ippei coded for grammatical knowledge, and filled the blank correctly. Informant Yuna coded for information earlier in the sentence (and as 'but' was circled on her manuscript it is plausible that this was that information) as well as for grammatical knowledge. She too filled the blank correctly.

In short, it seems fairly clear that in this instance the information provided by NCR and AC informants allows better insight into the processing of the item than can be

gleaned from think aloud data.

DELETION (13) [WERE]

GL1 think aloud Detlef

"*...slaves women and dishonoured persons <u>were</u> not allowed to compete...*"

???

GL1 think aloud Stella

"*...<u>were</u> not allowed to compete...*"

???

GL1 AC Bettina

Gra (WERE)

GL1 AC Heike

Gra (WERE)

JL1 think aloud Mitsuo & Arisa

M:".. this is *<u>were</u> not allowed to compete*

A: eh?

M: slaves and women were not citizens

A: *were not..*

M: *were not..* they couldn't attend..."

KOW, <u>Gra</u>

JL1 think aloud Yasuko

"*...slaves women and dishonoured persons <u>were</u> not allowed to compete ..'cause there should be a verb here and allowed is here and it should it's it's it's.. umm passive so I need here what d'you call..* in Japanese [copular] *like be and its past and its plural so it should be that should be were...*"

WC; Gra; LAN/L1Par

JL1 NCR Sachiko

"it's <u>were</u> here because there has to be a verb.."

Gra

JL1 NCR Hide

"I chose were because persons is plural and you need 'be' here"

Gra

JL1 AC Masa

Gra

JL1 AC Kazuko

Gra

Comments on deletion (13)

This deletion offers some interesting data, in that neither GL1 think aloud extract appears to offer any insight into recovery of the correct filler. This suggests that the filler was recovered with little or no conscious effort, as in think aloud such recoveries may go 'unexplained'. The first JL1 think aloud informants cite (historical) knowledge of the world, and the second offers a fairly detailed picture of exactly what grammatical knowledge she utilised. Post-task, Mitsuo was unsure whether he had made conscious use of his knowledge of English syntax, but felt the need for an auxiliary verb was "so clear". Mitsuo also noted that the extratextual knowledge he had used had entered his mind "at once". Arisa. could add nothing to Mitsuo's comments. Yasuko claimed post-task to have consciously called up her knowledge of TL grammar.

Both Sachiko and Hide again make explicit reference to grammatical constraints. GL1 and JL1 AC informants coded only for grammatical knowledge with no interpretable graphic markups or comments found on their manuscripts. There can be little doubt that, although grammatical knowledge appears to have sufficed to let AC informants fill the blank correctly, in this case the think aloud data provides the richer picture.

DELETION (18) [ALSO]

GL1 think aloud Claudia

"...*there was* mm...the problem is how to lump *boxing and wrestling* together [..] I could write *for example* but then you need a comma.. umm [QUERIES SINGLE WORD REQUIREMENT] *there was*.. hmm well..one word.. that's the problem 'cause no word fits there.. *there was*.. an article won't fit either.. there's already a verb.. if it was a list..if another noun followed then there would be a comma.. *there was* [...] hmm.. don't know it.. I'll go on to the next one..."

SWA?; LAN; Gra; LFN

GL1 think aloud Stephanie

"*there was <u>also</u> boxing and wrestling and dududu...*"

???

GL1 AC Rolf

SSf (ALWAYS)

GL1 AC Ollie

EP (ALSO)

JL1 think aloud Chisato & Kumiko

Ch: "*...there was also..* what?

Ku: there were this and that *and also* there were *boxing* and *wrestling*, right?

Ch: mmm.. it's continuing [GESTURES TO LINE ABOVE] right? *boxing and* something *and special test... varied ability...*"

EP (ALSO)

JL1 think aloud Masami

"umm.. this one.. I can think of various words.. always.. or also.. both are okay.. I'll take also.."

CBA (ALSO)

JL1 NCR Yoshiko

"It continues the list from before.. so also is maybe the best word.. but why is there a [new] paragraph here?

EP (ALSO)

JL1 NCR Miho **(ALSO)**

"I guess also is okay..or maybe moreover.. I put also"

CBA

JL1 AC Hiroyuki

No code entered; most of previous two lines underlined, with arrow to (18); <u>EP</u>? (ALSO)

JL1 AC Yumi

SSf(?)(EVEN)

<u>Comments on deletion (18)</u>

Although ultimately unsuccessful at filling the blank, Claudia appeared to look outside the text for a superordinate term to cover boxing and wrestling. Post-task she was unable to recall whether she had been thinking of a candidate phrase which would have fitted, had multi-word entries been allowed. She clarified her mention of

a comma, claiming that she had become aware that the blank could not be filled by a type of sport, as there would then have to have been a comma before 'boxing'. Claudia did not return to the deletion, and could not later recall whether that had been her intention. Stephanie claimed later to have filled the blank "at once" but added nothing more. Rolf's 'same sentence' coding may refer to the types of sport which follow the deletion. Ollie's coding implies that he was aware that the list of sports events continued in the new paragraph.

Chisato & Kumiko imply the same interpretation in their protocol, and this is underscored by Chisato's observed gesture. Even without the gesture, the remark that "it's continuing" along with the correct filler is probably sufficient evidence that the blank was filled with reference to the previous paragraph. Masami fills the blank appropriately, but provides only alternatives fillers withour rationalizing his choice. AC informant Hiroyuki used underlining, which was supposed to be used to indicate translated spans of passage, plus an arrow leading to deletion (18). This link to the previous paragraph was interpreted as a substitute for 'EP' rather than translation, and when approached after the session he was able to confirm that this was the coding he "should have" entered. One think aloud protocol in each L1 group seems useful here, and the other less so. NCR informant Yoshiko notes the continuation from above and queries the paragraph break. Miho, like Masami, offers alternative fillers but no explanation of her choice. The use of graphic markups on Hiroyuki's AC protocol offered a reasonable indication of the information the informant was utilizing from the previous poaragraph. (It is not easy to coherently summarise the information provided by AC informants' graphic markups of their manuscripts, although this information may be extremely valuable in some cases (see below.) (Gerloff (in Faerch & Kasper 1987) includes one perhaps not very generalisable attempt at representing such information graphically.) Yumi appeared to be referring to the types of sports following the blank, but her coding was not clearly legible as 'SSf' or 'SSb'. This, unfortunately, was not noted in time to query the point post-task. To sum up, think aloud and NCR seem to present a better picture of the processing of this deletion.

DELETION (20) [THE]

GL1 think aloud Detlef

"... *whose whose winner.. no the winner of which excelled as running.. I think I do not quite get this sentence but [...] hmm the winner of which excellent?.. is running also hmm hmm hmm the winner excellent is running and running [...] <u>the</u> winner of which.. whose winner...I think I'll skip that first and then go on.. it should be the best...*"

+Dif; CBA; <u>Sou</u>?; LFN

GL1 think aloud Alicia

"...excelled?.. *of which* no..not *of which* [...] *of which excelled of which* [SIGHS] well ... but it's about.. it's about.. *of which..*ah okay.. if that *excelled* is referring to one person then wouldn't it be *of which who excelled*? ... *special tests* requiring..f oh right..no..okay it's okay.. *determining..* it's referring to the pentathlon.. that's it! [I] was right all along.. ummm so it's *<u>the</u> winner* right?...."

Gra; Ah!; LAN; SSb; SSf

GL1 AC Rolf

Gra (THE)

GL1 AC Katrin

Gra (THE)

JL1 think aloud Chisato & Kumiko

Ch: "...*winner of which excelled in running* .. ha? something something wait wait.. *the winner of which..*the. pentathlon. right?.. *<u>the</u> winner of which* [...] *excelled in running...*"

SSb? (THE)

JL1 think aloud Hanako

pentathlon.. that's like triathlon? ... eh.. and *winner of which excelled in running* [...] umm *winner of which excelled in running...* surprisingly difficult!"

+Diff (--)

JL1 NCR Anzu

"I put down *<u>the</u> winner*.. but I'm not sure if it's all those events or like one big event.. if [the latter] then *each* could go in here too.."

SSf (THE)

JL1 NCR Sachiko

"you need an article here so the winner.. I also know the expression the something of something.."

Gra; Phr (THE)

JL1 AC Yuna

(--)

JL1 AC Emi

Gra; EP (EACH)

Comments on deletion (20)

Although this deletion was rated as relatively easy by both GL1 and Jl1 consultants, it seems to have caused more difficulty anticipated. Post-task, Detlef noted that although he had entered 'the', he had not been sure of his rejection of 'whose', even though 'whose winner of which' "sounded funny". He said that he had intended to return to this item, but had been too tired at the end of the task to do so. Alicia's protocol extract appears rather disjointed, but later she claimed that when she realised that only *one* person had excelled, she knew the filler must be 'the'. Asked if this was an instance more of the use of grammar, or of logic, Alicia answered that it involved both. Her claim to have been "right all along" suggests that she returned at some point to an earlier understanding, but she could elaborate on this post-task.

Both GL1 AC informants coded for grammatical knowledge, and both filled the blank correctly. No markups or comments were found on their manuscripts.

The first Jl1 think aloud informants' extract offers little insights into how the blank was appropriately filled, but post-task informant Chisato indicated that the realisation that 'of which' referred to the pentathlon showed that a single winner was involved, hence 'the'. Her partner mentioned that she had been thinking of 'every' as a filler, but felt that 'the' was better. Think aloud informant Hanako mentioned her surprise at the difficulty this deletion caused her, but could not elaborate post-task. Anzu seemed to rely on information contained later in the same sentence, but may also have been making a logical inference. Sachiko mentioned a grammatical rule, but also knowledge of a TL 'chunk'.

JL1 AC informant Yuna offered neither coding nor filler, while Emi coded for grammatical knowledge (but see below) and added a filler which was not in the SEMAC set. Again, verbal report data here was richer than the bare codings offered in AC.

DELETION (28) [WERE]

GL1 think aloud Anneke & Fred

Fr."*...all the victors..*

A. **..**victors <u>were</u> crowned yes.."

???

GL1 think aloud Stella

"*... on the fifth and last day all the victors <u>were</u> crowned with holy garlands of wild..*umm..."

???

GL1 AC Martin

Gra (WERE)

GL1 AC Lydia

Gra (WERE)

JL1 think aloud Masa

"*...and last day all the victors.. victors.. crowned..oukan*? [crown].. *victors.. and victors was.. <u>were</u>..* crown?...[L1 gloss of target construction] okay okay.. *were crowned with holy garlands of wild.. garlands?..*eh?"

Gra

JL1 think aloud Mitsuo & Arisa

F: *victors <u>were</u> crowned?...*

M: *were crowned..*yeah..*with.. garland...*"

??? (WERE)

JL1 NCR Hide

"well,, it just needs <u>were</u> here.. be crowned, right? I guess so.."

Gra (WERE)

JL1 NCR Yuko

"hmm.. do you say get crowned?.. be crowned? [repeats alternatives] sounds okay maybe.."

CBA; Sou (GOT)

JL1 AC Ippei

Tr; Gra (WERE)

JL1 AC Risa

Gra (WERE)

Comments on deletion (28)

The protocol of Anneke & Fred offers no evidence as to how they recovered the filler, showing only that they did so very quickly and in full agreement. Post-task, Anneke indicated that she had "just known" that an auxiliary verb had to go there, and that it must be plural. Fred added that he had recognized it as a passive construction, and that he had already begun to render the clause into his L1 because he did not know the upcoming word 'garlands'. Stella also filled the deletion 'in passing', without interruption to her reading aloud of the passage span. Both GL1 AC informants coded with 'Gra'. Masa used an L1 passive construction in glossing the phrase 'were crowned', which suggests that she had recognised the TL structure as passive. She confirmed this post-task, adding that she had momentarily taken 'garlands' to mean 'those doing the crowning.' Mitsuo & Arisa. also later noted that they had recognized this immediately as a passive form, with which they were familiar from school English lessons. Hide cited grammatical constraints here, but Yuko chose the wrong one out of two alternatives apparently based on which sounded best. Again, both JL1 AC informants coded for the use of grammatical knowledge.

In this deletion, then, three out of four think aloud protocols give no real insight into processing, while all AC mnanuscripts tell us that informants were making use of their knowledge of TL grammar. The Jl1 NCR data seems to fall between the two. AC informant Ippei's coding of 'Tr' was queried just at the end of his AC session, and his comments indicated that he had translated the clause in order to "check" his understanding of it and to try to work out the meaning of 'garlands'. (Ippei had been a low-verbaliser in an earlier think aloud session.) Asked whether he had translated before or after writing in the filler, he thought that he had done so before. This suggests that translation played at least a secondary role in Ippei's recovery of the item (perhaps with 'Gra' as a confirmatory cue?) even though at least part of his focus was on an unfamiliar extant passage item. In summary, verbal report here does not appear to tell us much more than could be gained from AC codings.

DELETION (2) [LOST]

GL1 think aloud Claudia

"...*eventually* mm the connection is missing umm.. I think a verb could come in there relating to Zeus but I'm not sure.. umm.. I'll go onto the next one..."

WC; LFN; <u>SWA</u>?

GL1 think aloud Lise

"*... eventually local character became <u>lost</u> local character became first...*"

???

GL1 AC Walter

WC (--)

GL1 AC Andy

WC (LOST)

JL1 think aloud

"*...eventually..*ah.. verb right? umm (6 secs) *<u>became</u> its local character.. local character.. became first...*"

WC (BECAME)

JL1 think aloud Yoshiko

"*... I guess it.. It's a verb* . it *lost its local character I guess* [...] *first it lost local character and then it became.. national then international...*"

WC; SSf; <u>Log?</u> (LOST)

JL1 NCR Jun

"I put in changed its local character because it says here that it became national and then.. international.. so it changed a lot.."

SSf (CHANGED)

JL1 NCR Ayumi

"I couldn't think of a good word to go in here.. I know it must be a verb and it means that the Olympic Games became very different.. [...] I thought of exchanged but then you must have .. exchange something for something. I'll come back to this one.."

+Diff; WC, TLParaph; LFN (--)

JL1 AC Tomomi

WC (--)

JL1 AC Kazuko

WC; SSf; (WAIVED)

<u>Comments on deletion (2)</u>

Claudia realises that a verb is required, but cannot fill the blank and moves on.

Post-task she claimed to have been considered 'gave up', but could not think of a single-word equivalent. Lise offers no insight, and later remarked only that the recovery happened "automatically". Both AC informants coded only for word-class. 'Verb' was written in English in the comment box on Andy's manuscript.
JL1 think aloud informant Yoshiko realised that a verb is required, but chooses an inappropriate one. She also noted that the filler must be a verb, and recovered deletions (2) and (3) together. Post-task, she noted that the games had to lose one quality to gain another, and confirmed that she felt this was a logical inference. Jun uses information from the same sentence, whereas Ayumi gets the class of the missing item but then moves on, not to return. Again, the JL1 AC informants coded for awareness of the filler's word-class although Tomomi left the item unfilled. As her manuscript clearly shows (a line links the passage item 'waiver' to the deletion) Kazuko noted the verb 'waived' later in the sentence and applied it in deletion (2). She also accurately coded for this event, and the sequence of codings suggests that she ralised a verb was needed and either deliberately or by chance found one which was unfortunately not correct. Lise and Jun both recovered the item successfully, so that they clearly 'knew' its word class. All the others at least recorded some awareness of this. The prevalence of 'word class' coding here indicates that informants in all conditions had difficulty with it. The richer picture seems to be from verbal report data in this item, with some evidence of where informants Yoshiko, Jun and Ayumi sought cues. There does not appear to be much to choose between data gleaned through think aloud and that from NCR. Kazuko coded for 'same sentence forward' information, but other AC manuscripts are bare of unambiguous further data.

DELETION (7) [RECORDS]

GL1 think aloud/1 Detlef

".. *olympic games go but some official* ehmm *some official* ehmm guesses (*Schätzungen*)..what is guesses..*translated*..guesses [...] umm *exp.. no expectations is wrong but* ehmm *guess...the guesses no but the noun of guesses.. I don't know right now but I think it should be something like that so even if it is wrong put in guess..*es.. *I think that's quite German but anyway...*"

L1Eq/Tr?; CL

GL1 think aloud Tom

"...*but some official*...umm yeah umm and then it gives the date what can it be? umm..think of...*the date* or.. or *state* umm *state.. the date* or something like that?..*date from*..hmm.. from *seven and seventy six*.. okay so it gives more precise dates [5 secs] *some official dates?*..yeah.. it's about when they took place..."

SSf (--)

GL1 AC Janna

Col; L1Eq (INFORMATION; margin comment '*Infos*')

GL1 AC Bettina

Phr; Tr/L1Eq (INSCRIPTIONS [sic]; margin comment '*Inschriften?*' *plus small image*)

JL1 think aloud Kiyoko

"...*but some official official.. some date.. official date? official beginning?.. date official*..(CLICKS FINGERS) *starting date..date from...*"

Sou?; Col? (STARTING)

JL1 think aloud Chisato & Kumiko [Note: Ch.&Ku. appear to be revising deletion (6) [THE] while also trying to fill deletion (7)]

Ch: *some official...date from..* Ku: ah.. *sakanoburu!* (go back in time) *date back to..go.. some official official* so isn't it *game*? *games*.. or..

Ku: no.. eh.. [...]..it's a verb..

Ch: ah that's *date back to* [...] ok ah ok *some official...*

Ku: so they mean *game*, right? *games*... yeah so that noun comes here

Ch: *official..*

Ku: *taikai* means game, right? Ch: *date from..* yeah.. so small letter is ok [...] if you don't use something like has.. *olympic games go back* to?

Ch: as.. as *olympic games go..* it..ah wrong.. no.. it doesn't sound right.. [...]

Ku: *how far back..* anyway it means they don't know since when they've been doing it..

Ch: yeah.. but official games is based on this, right? {POINTS TO DATE '776**BC'**]

Ku: eh?

Ch: how far back..

Ku: oh, isn't [deletion (6)] *unofficial*?

Ch: ah, then [deletion (6)] contrasts with *some official*? Ku: yeah..

Ch: ok with *go back* completes this here..

Ku: yeah that's right [Ku. ENTERS '*GAMES*' IN DELETION (7): Ch. NODS]

Tr; WC; Sou; SSb; SSf

JL1 NCR Miho

"I put down official records.. I know those [words] go together because I had to send an official record of my studies to [U.S. college].. like my transcript.. so I know this expression.."

Phr/Col (RECORDS)

JL1 NCR Anzu

"I don't know this one. I could only think of official papers but [notes that paper had not been invented at that time.]"

KOW (--)

JL1 AC Yuna

Phr (DOCUMENTS)

JL1 AC Emi

(MATTERS)

Comments on deletion (7)

Detlef's protocol indicates that he is uncertain of his filler, and post-task he claimed only to have entered it because he could not think of anything better. (He also mentioned at this point that he had felt under some time pressure owing to the fact that he had arrived late for the appointed session, and hence felt unable to spend very long thinking of a better filler item.) He was not able to say how much of the passage context he had in fact translated, but could confirm that he had glossed at least that clause in his L1. With hindsight, he felt that 'official estimates' would "go together better", which in terms of register is correct. Tom confirmed post-task that he had referred to the following sentence context, i.e. the historical date, but rejected 'dates' as a filler with the implication that he felt he might be being led astray by the L1 item *Daten* (data, facts.)

Clear evidence of a role for the L1 is also found in both GL1 AC informants' manuscripts, one of which contains a vernacular abbreviation for *Informationen* and the other an L1 equivalent of the chosen filler. These informants also code for awareness of phrasal or collocational links

Post-task, JL1 think aloud informant Kiyoko confirmed that she had evaluated candidate fillers by whether or not they sounded right, or at least sounded like words that "go together". She noted that had recalled the pairing 'starting date' and could think of no better filler at the time. Asked if she had understood 'date' in 'date from 776BC' as a noun, she indicated that she had. Informants Ch. & Ku.'s protocol shows that they discussed this deletion at some length, again with a role for L1 glossing. Chisato felt that she had probably translated quite extensively, although her L1 paraphrase of the passage context (in the PTI) was not entirely coherent. Kumiko felt that she had "thought in English" except for seeking L1 equivalents of a few items. These two informants referred back to the context of the previous deletion, and in fact revised their filler for deletion (6) [unofficial] apparently in the light of their processing of (7). Miho seems to have applied knowledge of a collocational relationship arising from a recent real-world experience, while Anzu rejects a candidate filler on the basis of extratextual knowledge. Yuna coded for phrase-level knowledge and provided a SEMAC filler, while Emi filled the blank inappropriately and offered no coding at all.

For this deletion, then, the insight provided by think aloud and NCR (both reasonably full) exceeds that gleaned from AC, although part of the information which think aloud provided stemmed from post-task interviews.

DELETION (8) [TOOK]

GL1 think aloud Stephanie

"...*take* aah.. *the games took place* because it happened in the past..."

???

GL1 think aloud Claudia

"...mm *the games*..[/pleiz/] *in August on the plain* umm..I'd say something like *always* could come here.. or.. yeah..actually *always* because.. *in August*.. it's to do with the repetition [of the event] each year and so there has to be a kind of time-related word to complete the sense.. *always* actually wouldn't be bad in fact..."

SSf; Log; Tr

GL1 AC Vince

Phr (TOOK)

GL1 AC Ollie

Gra (TOOK)

JL1 think aloud Kiyoko

"*... the games took place.. take hold.. took.. the games tooks.. took place in august...*"

Sou? (TOOK)

JL1 think aloud Yumi

"*...that games.. games* mmm *held* mm? *the games were held* eh? wrong?.. no wait.. *games games..* something or other *place.. games had.. had a place in august* [SIGHS] [5 secs] *take place maybe.. took place the games took place in august...*"

Sou? (TOOK)

JL1 NCR Toru

"I put take here.. I know the expression to take place.. "

Phr/Col (TOOK)

JL1 NCR Hide

"took place.."

??? (TOOK)

JL1 AC Tsuneo

Phr (TAKE)

JL1 AC An

Phr (TOOK)

Comments on deletion (8)

Stephanie's protocol gives nothing away beyond the fact, confirmed post-task, that she had recalled the phrasal verb in its dictionary form, and at once recast in the appropriate tense. Claudia's protocol is much more interesting here, for her (mis)pronunciation of 'place' suggests that she had interpreted it as a verb. She had correctly grasped from the following 'in August' that a recurrent event was referred to, and from this she logically, if wrongly, concluded that 'always' would be a suitable filler. Post-task, Claudia seemed surprised to learn that her filler was not appropriate, but (having asked that I not tell her the correct answer until she could try once more) she quickly realised her error and supplied 'took'. She confirmed my suspicion that she had taken 'place' to be a verb, and thought that she had glossed or grasped this construction in her L1 as '*die Spiele spielten sich in August ab*' (the games took took place in August.) Claudia was clearly annoyed with herself over this error, as (she

said) the phrasal verb 'to take place' was "quite familiar" to her. This is an instance in which think aloud clearly performs better than AC, although GL1 AC informants' codings reveal that Vince interpreted the connection of 'took' and 'place' to be a phrasal one (On Vince's manuscript the filler took+place are enclosed in a box.) and Ollie a grammatical one. Given the phrasal verb status of the 'chunk', this is perhaps not surprising. (On Vince's manuscript the filler took+place are enclosed in a box.) Kiyoko appears to have 'sounded out' at least one alternative to her initial (and correct) candidate filler. Post task, she indicated that she had not been sure that 'took place' was appropriate, and had indeed sounded out alternatives, of which she could recall none until she audited her recording. Informant Yumi also put obvious effort into filling this deletion, and she too confirmed that she had begun sounding out possible fillers as soon as she realised (having recalled the single-word requirement) that her initial choice of 'held' was not suitable as it required an auxiliary verb. Once again, think aloud clearly offered more insight than the bare AC codings for use of phrase-level knowledge offered by Tsuneo and An. The products of NCR are **mixed**: Toru's protocol reveals his knowledge of a phrase or collocation, while Hide supplied the correct filler but gave no indication as to how.

DELETION (14) [SEQUENCE]

GL1 think aloud Alicia

"... well.. *the exact dates* I think *of events is uncertain*... oh no they.. I see that later on they're described in more detail..so.. ah..*the exact sorts of or kinds of events* ..hmm.. maybe *kinds* is better but I'm not sure..."

SSf; CBA

GL1 think aloud Julia

"... *the exact* well.. sequence.. *turn of events.. number of events* probably *number* because it isn't about what they are so much as the fact that they take place.. **so** *events*.. yeah and then some are listed..."

Sou;Log;SSf (NUMBER)

GL1 AC Bernhard

Col (NUMBER)

GL1 AC Kristen

Phr; L1Eq (SEQUENCE)

JL1 think aloud Yasuko

"... um *the exact...events is uncertain but events* dadada.. *dates.. no this is not date is that* [...] *of events* ? umm.. *exact course of events.. no it cannot be course of events 'cause there are other* umm.. *kinds of events following this sentence.. so there should be number the exact number of events is uncertain but events ...*"

Cha; SSf

JL1 think aloud Rumi & Satoko

Ru: "*...the exact 'number'*, isn't it?

Sa: *number*? number?

Ru: *number of events*.. that's not certain, but... hmmm"

??? (NUMBER)

JL1 NCR Jun

"I put plan.. the exact plan because there must have been starting times and so on.."

??? (PLAN)

JL1 NCR Anzu

"I put number because I know you can say that.. like the exact number of something.."

Phr/Col (NUMBER)

JL1 AC Risa

Phr; SSf (NAMES)

JL1 AC Tsuneo

SSf (NUMBER)

Comments on deletion (14)

Alicia's protocol indicates that her rejection of her initial filler is based on information subsequently identified later in the sentence, Her candidate fillers are both plural, however, which conflicts with the following verb phrase. Post-task, Alicia. seemed surprised when this was pointed out, and she expressed some uncertainty about subject-verb agreement in cases such as this: could one say 'kinds or sorts of events', or did 'event' have to be singular? She then proposed the L1 equivalent *Reihenfolge*, which she glossed as (SEMAC) 'order'. Alicia did not think that this L1 item had occurred to her while processing the deletion, however. Julia recovered the L1 equivalent of the original filler word as Alicia did, but chose

'number', thinking, as she suggested post-task, that it "sounded better" than 'turn'. Sound alone, however, does not seem to account for what looks like a logical inference in Julia's think aloud, and post-task she agreed that some kind of inference had been made. AC informant Bernhard coded with 'Col', with an arrow leading from the list of events to the deletion. Kristen chose 'Phr', and also noted the L1 equivalent *Reihenfolge* in the comments box. Both informants filled the blank appropriately. Yasuko revised her initial filler 'course' on the grounds that other kinds follow. Post-task, Yasuko could not clarify this decision, except to say that 'course', for her, implied a list of types. As events were listed later on, 'number' seemed the better filler for (14.) Pair-member Rumi could not offer any further information about how she arrived at 'number', although she claimed to have been "not very sure" that it was correct. Satoko claimed that she had been dissatisfied with the filler, but had not been able to think of a better one. NCR informant Jun's proposed filler did not appear on the list of SEMAC alternatives for this item (although if not right, it has something right *about* it) and he may have been led astray by the Japanese *yotei*, which can mean either plan or schedule according to the context of use. Anzu coded for phrasal or collocational knowledge, which is amply supported by her protocol. AC informant Risa coded with knowledge of a phrase that does not exist, while the circling of 'racing, field events' with an arrow pointing to deletion (14) indicates that she also attended to the sentence context following the blank. This is not reflected in her own coding—a point which I take up below—and so my addition of the coding 'SSf' is shown above in underline. All in all, for this item the better picture is provided by verbal report.

DELETION (26) [ASIDE]

GL1 think aloud Stephanie

"*...full moon was set up as a holiday* ..that'll certainly do *..set up* [...] *was set up as a holy day ...*"

???

GL1 think aloud Tom

"*...of the full moon was set* ..emm aah well.. I'd guess it was in any case...as a holiday..this day.. emm this moon was [taken?] as a holiday (5 secs) *was set officially* perhaps? *as a holy day* or *was set finally* [...] ah.. no this isn't yet the end of the

games.. but it was fixed [*festgelegt*] anyway.. *was set officially as a holy day* or something like that.. that seems more natural..."

Sou; Log

GL1 AC Rolf

Phr (UP)

GL1 AC Lydia

Col (OFFICIALLY)

JL1 think aloud Kazue

"...*of the full moon was set set set.. was set* something *as a holiday.. was set set set set.. as set as as* something *as was set* mmm? *was set* ah *was set holiday as a holiday on the* okay it was..something.. *set* ah *set* ummm *set set set full moon was set off? set on as a holiday?..*"

Sou? (OFF)

JL1 think aloud Yasuko

"*and the fourth day* something umm.. set? as a holiday.. *fourth day of the full moon..* of the what? I can't think anything that comes here [INAUDIBLE] *the fourth day of the full moon*? *was set* (7secs) *fourth day* something *of the full moo n*l dunno.. f*ourth day* something *of the full moon was* [...] f*ull moon.. fourth day was set* something a*s a holiday set aside*? hmmm? *holiday* to be [A HOLIDAY].. I don't know about this here hmm.. *set* something it *[I] think it means like it was like kept as a holiday.. or like it was set as a holiday or something I.. set* umm.. something was a holiday think.. *every fourth day of the full moon was..* there wasn't any event but then what shall I say?.. was set aside as a holiday [...] I do it later."

TLParaph; CBA; +Diff; UND; LFN (--)

JL1 NCR Toru

"I wrote in was set off *as a holiday* but I don't think it's correct.. I think I understand the meaning.. it's like *totte oku* (reserved) in Japanese.. I can't think of another expression.."

UND; L1Eq (OFF)

JL1 NCR Ayumi

"I couldn't think of a word for this one.. I'll come back.."

LFN (--)

JL1 AC Etsuko

Phr (UP)

JL1 AC Atsushi

Phr (--)

Comments on deletion (26)

Stephanie appeared confident in her choice of 'set up' as a filler, and post-task she held quite firmly to the view that it "should be" a SEMAC filler. Tom appeared to be sounding out candidate fillers, with an apparent logical inference telling him that 'finally' was not appropriate here as the games had not ended. Post-task Tom claimed to have selected 'officially' because it "sound[ed] better". AC informant Rolf coded for use of knowledge of phrase, but also selected 'UP' as a filler. Lydia's code appears to indicate an assumed collocational relationship between 'set' and 'officially' (although a corpus search reveals that the 'unmarked' order would in fact be 'officially set'.)

JL1 think aloud informant Kazue appears to rely on sounding out, but fails to fill the blank appropriately, while Yasuko seems to have found then rejected the appropriate 'set aside'. She paraphrases the meaning of the span accurately in the TL, but in the she abandons the item. The fortuitous pairing of these two extracts is, like the protocols of Yasuko and Harumi in chapter 6, a good illustration of the range of productivity think aloud informants can display. Post-task, Yasuko claimed to have had trouble deciding between 'aside' and 'apart' as her choice of filler. JL1 AC informants Etsuko and Atsushi also coded for phrase-level knowledge, but neither filled the blank appropriately. Here, too, verbal report offers a better picture of informants' task processing behaviour.

Addendum: a note on GL1 AC data-collection

Given that I had gathered data from German and Japanese L1 think aloud informants, it seemed worthwhile to collect AC manuscripts from German informants, too, when the opportunity arose. The time available to gather this data–using16 German L1 informants recruited from participants in an intensive ESL course at a central German university–was very limited, and barely sufficed for the orientation session and the AC cloze task, with no time for the administration of comparison tasks as had previously been done.

These informants' course was led by a former colleague with personal knowledge of my earlier GL1 TA informants, however, and so it was possible to establish through informal assessment that those two informant samples were comparable in terms of target language ability. A t-test subsequently applied to GL1 AC scores and JL1 AC scores revealed no significant difference between means (-1.303 ns. at 0.05 df38), suggesting that the task was within the ability range of both AC informant groups, and that the rather shorter orientation session possible with GL1 AC informants either (a) did not significantly affect their performance on the cloze task (My orientation of the GL1 AC group, conducted in German and English. was arguably more concise than the same orientation conducted with JL1 AC informants.) or (b) was compensated for by a slightly higher target language proficiency on the part of the GL1 informants.

Jl1 and Gl1 AC results were briefly compared in chapters 7 and 8, but it may be worth adding here that as the AC session time was limited for the GL1 group it is impossible to say whether the noticeably lower quantity of 'additional' information (graphic markings; written comment, etc.) found on their manuscripts reflects merely the time constraint, or a real difference in approach to the task.

Appendix 5: PCTA Protocol of JL1 informants
Mitsuo & Arisa

NB: TL verbalisations are shown in italics, and all other speech translated and shown in regular type. L1 speech is retained and shown in italics where minor differences in interpretations were thought possible. Researcher comments and codings are shown in brackets.

A: *four*? [KOW]

M: *four.. four years..*

A: let's write it

M: yeah.. write it

A: [READS INAUDIBLY]

M: let's read on.. we don't know what's written here.. [Rpt]

A: [LAUGHS]

M: what's that '*official*'? a document? [Phr/Col/Idi] [CP]

A: I don't know

M: *official document* goes back from 1976 to..
A.
mm..
M.
perhaps it says when it started..
[READS AT LOW VOLUME, THEN APPEARS TO TRANSLATE] [...] *no* one knows' is right..*no*? [Phr/Col/Idi]
A:
mmm..

M: *nobody knows*.. but.. seven hundred and sixty years..

A: uh.. [LAUGHS]

M: perhaps, because it's '*some*' [Diff?]

A: [WRITES] *no one knows*..

M: something comes here.. but what? [Phr/Col/Idi]

A: the olympic games

M,A : [LAUGH]

A: let's write in pencil what we think.. later on..

M: let's make this one '*no*' [CP]

..

M.

something comes here, but what? [CP]
A.
the olympic games
[M., A. LAUGH]

A. docu..
M.
DOCUMENT (*kiroku*) so... record? document? [A. FAILS TO RESPOND TO M's QUESTION] ah... *took* place in August.. is right
A.
mm.. [READS AT LOW VOLUME]
M: mount.. mount.. ah there's mount olympus

A.
mm?
M.
plain..
A.
[LAUGHS] *of*?... don't know..[GS?]
M.
shall we make it *of*? mount olympus.. no.. *took* place.. [CP]
A.
[WRITES] took place [INAUDIBLE]

this is *were* not allowed to compete.. [Gra]
A.
eh?
M.
slaves and women were not citizens.. [KOW]

A.
were not..
M.

were not.. they couldn't [even] attend.

A: couldn't come.. but

M: even that something something what.. *gaiya* [baseball outfield] [TR]

A: ah

M: *outfield* is something like *gaiya*, no? [CP]

A: mm...

M: is it '*came*' here? [CP]

A: mm.. *no married woman were allowed*

M: check?

A: was..

M: I see let's leave it for now / make it horyu [let's postpone it] [LFN]

A: postpone [LAUGHS]

A: *the exact..*

M: [INAUDIBLE]

A: *the exact number..*

M: is it *number*? isn't it exact purpose they don't know? [Diff?]

A: is it event? *exact..*

M: *seikaku na* [exact].. *shumoku* [event] or *mokuteki* [goal]? [TR]

A: e*xact number of even*t.. *include* to ka [and so on] here?

M: *include.. included..* [WC?] {Sou?]

A: *include.. boys' gym..*

M: *horse racing..* aren't these events? [LOG? KOW?]

A: how can I say this [events]?

A, M: [INAUDIBLE]

M: *pass* [let's do it later] [LAUGHS] [LFN]

A: *field events such as to ka* [LAUGHS]

M: this is example no? *foot race..* all these are examples explaining *field events* so *such as discus..* such an event must have existed, whatever they were [LOG] [CP]

A: what's *discus*? [UPI]

M: whatever.. *javelin.. javelin throwing..* I don't know what it is but it must have existed and.. [LOG?]

A: *and? also?* [CP]

M: what could it be? [CP]

A: umm.. I don't know these words [UPI]

M: words [*tango ga*]

A: take that [...] what is this? [CP]

M: how are these connected? there was.. [CP]

A: huh? *there was..* skip it? [*tobasu*] [LFN/GUP]

M: but isn't it here in order [*junban*]? after *excelled in running, discus and javelin* appeared before.. we didn't know then.. *kore da* [this one]! [SPb]

A: and

M: *and wrestling* [SIGHS, LAUGHS]

A: let's read it through once [REV]

M: mm

[READ INAUDIBLY]

A: isn't it '*of*'? check.. [LAUGHS] [???]

M: as the previous spaces are empty I'm uncertain..

A: let's read the previous parts again [REV/EP]

M: yeah.. let's.. but we may get stuck in the middle

A: *eventually* [INAUDIBLE] oh ..no no

M: as the event was held [*okonawarete*] *became* something

A: *became first..*

M: here it is.. subject.. so 'of'? of *foreign competitors have been waived* [/waIvd/].. is '*of*' correct? [GRA], [

A: what's this? [CP]

M: this I don't understand.. [/waIvd/] [UPI]

A: [LAUGHS]

M: this must be '*of*'..*of foreign competitors* waived [/waIvd/] [READS INAUDIBLY] back to? no one knows what is it? what could be put in here?.. go back.. there's go..

A: not 'the'? [Diff]

M: yeah 'the'.. I thought so too.. that's right here it says *the olympic athletic festival*.. that's right.. let's put '*the*'.. that's good isn't it? this one is right I think I don't know which though..*document*.. I think *official document goes back to 776BC* [EP]

A: check?

M: plain..if not..impossible [IT WOULD BE IMPOSSIBLE IT IT WERE NOT HELD ON THE FIELD OR PLAIN] [KOW/LOG]

A: [READS INAUDIBLY] ah..invaders.. are?

M: what? but why is it 'was' here?.. *but married woman was.. admitted.. spectator* these are all singular what is this? this I don't understand why is it singular here?.. *many thousand spectators.. [ooku no kankyaku ga].. came from Greece but..* it requires historical knowledge [TR], [KOW]

A: isn't it because it's '*no*'? [DIFF]

M: *no one.. no..* [*daremo-nai*] ah.. I see I see

A: *no married..* then here as like (as) '*a*' [GRA]

M: *as a..* yeah right! then it's comprehensible so married woman are not allowed

A: right right

M: up to here it's okay

A: the exact '*number*', isn't it? [CP]

A: *number of events..* that's not certain, but..

M: but.. *boy's gymnastics..* what is it? *boy's* what? strength [*tairyoku*] [KOW], [TR]

A: GYMNASTICS [*taiso*]

M: gymnastics? *horse riding such as..* these are field events, these discus and javelin throwing and..

A: and.. *field events such as..* so.. *field events are such as* this and this and the last one's this so the order of examples of events are *boy's gym* is one two three is *field events..* and the last one with.. and four *include include..*[LOG]

M: must be it *include.. including* is better? [WC?]

A: isn't it '*also*'? [Diff]

M: yeah there are four examples.. is it a number here? there was..what? [READS INAUDIBLY] *there was boxing and wrestling* or was there? wasn't there?

M: wasn't there? [CP]

A: there may not be now.. but..

M: what's this? it existed or not? [CP]

A: what's '*test*'? special tests? [CP]

M: what's this? check up [*kensa*]? [CP]

A: ah I see

M: *special tests*?

A: *special tests..*something *varied ability* .. I don't know this [UPI?]

M: what's this? difficult!..[+Diff]

A: *pentathlon*..

M: using which..*winner which excelled in running*? what could it be? is it okay to skip it? [LFN/GUP] [CP]

A: skip it

M: skip it? [CP]

A: mm..

M: okay, skip it

A: next! preposition [*zenchishi*] isn't it [WC] [CP]

M: preposition.. perhaps 'at' but we need an article [*kanshi*] as it must be '*the evening*' of.. what could it be? [GRA]

A: *sacrificial*? I don't know..[UPI]

M: sacrifice [vb: *gisei ni suru*] [TR]

A: 'sacrifice'?

M: yeah.. may be sacrifice..

A: [READS INAUDIBLY].. *heroes.. of the day*. [???].

M: *of the day*..

A: as this is the third day.. [LOG]

M: yeah and the fourth day.. [SSf]

A: well, if the preposition comes here, '*the*' is still needed.. [GRA]

M: right right what is it? to? what does it want to say? heroes? [CP]

A: on the third day a scrificial offering [*osonae mono*] is.. [TR]

M: to the hero(es) of the day? [CP]

A: to the hero(es).. offered to the hero(es) on the fourth day

M: *on the fourth day*.. full moon night? [TR/L1Eq] *mangetsu no hi*]

A: ah.. *night of the*.. [Ah!]

M: yeah *night of the*..

A: *was set*..

M: *yo-ho usage* [how to use]? [LAN]

A: *set*..

M: maybe [...] eh? holy day [*shinsei na hi*] what? holy day [...] [TR]

A: something *and last day*..

M: *third and last day*? [CP]

A: ah..

M: last day and the day before last, aren't they? victors.. victory.. *shourisha* (victor)? [CP]

A: *victors were crowned*?

M: *were crowned*..eh..*with*.. *garland*

A: [INAUDIBLE]

M: *wild*.. sacred wood..from..

A: ah! that! the leaves.. like that crown of leaves.. what was it? [UND] [CP]

M: yeah! what is that kind of hat called? [CP]

A: *garland of wild*..

M: this one's right with '*were*', no? [???]

A: mm..

M: *crowned with garlands of wild*.. *of wild*

A: isn't *garland* a hat? [CP], [KOW] [UPI]

M: hat.. and hat of what.. of wild something.. *yasei no* [wild] something.. *shinsei na ki* [sacred wood] [TR]

A: *wild*..

M: *wild*.. *sacred wood*.. *wild*.. *miki* [trunk]? what's *miki*? trunk? no, that's strange..[TR]

A: *wild*.. *from a sacred wood*.. *toreta* [produced]? [LOG?] [CP]

M: produced from sacred wood.. some thing wild

A: *wild*.. what's *wild*? *yasei*? [TR] [CP]

M: *yasei teki*

A: *leaf*? [CP]

M: *leaf*? that's good! *wild leaf*.. so.. made of wild leaves

A: produced from sacred wood.. you know, those green leaves made in this way [GESTURES] leaves? [UND]

M: *leaves*

A: *so great was the honour.. that the winner of the footrace..*

M: which is the subject? [GRA] [CP]

A: I don't know

M: *foot race*.. what goes here? [CP]

A: this? this is a verb, no? *so great*.. [WC]

M: yeah, right.. *the honour that*.. is *doukaku* [juxtaposition]? the honour [*eiyo*] was great.. no? that's right? [GRA]

A: *amari ni mo.. nano de* [so _____ that] *that the winner*

M: oh, so this is '*so* ____ *that*'..

A: yeah.. so___ that

M: yeah? but if it's a '*so*___ *that*' structure isn't the 'so' part too complicated? [OFFERS ILLUSTRATIVE EXAMPLE] *kare ha hijo ni tsukare te ita node hatarake nakatta* [he was so tired that he couldn't work] '*so*___ *that*' is this kind of structure, isn't it?

A: *so great was the honour that..*

M: *touchi* [inversion]? it's better to think of it as inversion? [GRA]

A: eh? inversion? I don't know what that is

M: I'm only putting it in terms of grammar

A: I don't know grammatical terms [IRRITATION?] *so great was..* the a*marini mo eiyo ga sago kattano de winner wa nanika ni namae o ageta* [honour that the winner gave his name to something] *name to year of his victory*? [TR]

M: *foot race* appeared before.. where? *very important foot race*.. here it is..

A: but what does it mean.. *gave his name to*.. [CP]

M: his own name to.. what? the olympic year? [CP]

A: what is the year of his victory? *katta toshi*? [CP] [TR]

M: *katta toshi*.. if Mr. Something won, then it became Mr. Something's game [*taikai*].. I don't know..

A: *winner of*.. only one word goes in the space? [SWA?]

M: right.. *the winner of*.. *important foot race*.. ah 'important foot race' must be *dash-kyou sou*.. 100 metre dash etc. Carl Lewis, for example.. if he won, it became Carl Lewis' year.. the winner of 'the'? [KOW]

A: *winner of important*?.. but then we still need 'the' [GRA]

M: year.. 'an' important or something is needed.. but only one word [in] each [space].. gave his name.. gave A to B.. *gave his name*..

A: eh.. mm

M: a bit difficult, but it became clearer..

A: let's go back [REV]

M: yeah, we have time.. let's go back

A: *eventually* [...] *its local character*..

M: what's *local character*? ah.. ah I see.. Zeus is the highest god of twelve Olympus gods.. there are other gods as well.. Venus Appollo etc. I remember learning them before.. *eventually*.. *eventually*.. Zeus *eventually*.. what? '*of*' is right, no? *of its local character*.. is 'local god' [*tochi-shin*] no? [CP] [KOW], [TR]

A: feels that way..

M: yeah right? finally [*saishuteki ni wa*] became local gods and then became first.. then.. *international* [SSb/SSf]

A: yeah, first something [*nantoka*] and then

M: then..

A: then this *after the rules*..

M: first is one and then is two.. *after the rules*

A: [the] next [one] is '*international*'

M: okay.. what comes before *international*.. eh.. [Phr/Col/Idi]

A: first (a) *local event*.. '*local*' appeared somewhere

M: yeah..it appeared.. here it is..[SSb]

A: I see

M: they started to worship a local character [i.e. A LOCAL GOD] then it became *a local event* the something.. international one international [*kokusaiteki*] or within Greece it became *international.. international..* I mean among *polis* in Greece [KOW]

A: mmm..

M: maybe.. must be!

A: yeah, like that

M: like that.. making it up [*dechi-age*].. I don't know.. then this here is local [INAUDIBLE].. *foreign competitors*.. I see I see..foreign [*gaikoku no*] competitors.. ..foreign here means from other *polis*.. competitors came.. rule became international.. 'waive' means something like 'accepted'? 'expended'?

A: no..[DIFF]

M: [READS INAUDIBLY] umm..

A: umm.. it's okay around here

M: but on the record.. um..okay 'knows' is *genzai* [now].. the date is okay.. make it '*records*'? because it's 'some'.. make it 's'.. '*records*' or '*documents*'.. um.. the game [READS INAUDIBLY] okay here, right? but [READS INAUDIBLY] so okay, *married women were not allowed* and then.. *women..slaves.. dishonoured persons..* this I don't understand, but it doesn't matter.. okay.. conditions of particpation [*sanka-shikaku*] is down here.. *the exact..* the number is not certain, but the games had the following events.. then four events are listed.. right? horse racing is *jouba*? [horse racing].. *jouba.. de..* running to the other side etc. right? field events.. *discus and javelin throwing..* these are field events, right? they appeared later too.. *discus and javelin throwing..* shot putt [*hougan-nage*]? {SPb/SPf]

A: ah shot putt!

M: throwing to the other side..

A: umm..

M: I heard that in the old days they decided the area of the property one possessed through throwing a putt [KOW]

A: mmm..

M: so, it's shot-putt.. they must have done it.. and very important.. *de..* this is short-distance dash/sprint [*tokyou-so*].. there was there's some new information here.. that means there wasn't boxing and wrestling? but 'no' would be in here [LOG?]

A: but why (have) tests? [CP]

M: yeah, yeah..

A: why does it have to say [mean?] there are no tests? [CP]

M: hmm.. what's the function of 'and' here?.. it's not clear.. various [*samazama na*] ability?.. such as the.. this is the abilities which follow.. for example there was (ability in) pentathlon.. such as the pentathlon.. why? why use 'such as'? afrer such as, there are several examples.. I see I see [READS INAUDIBLY] what is it? but there's '*wrestling*' here, so there was [EMPHATIC] wrestling..

M: there were.. there..

A: also? [CP]

M: *tests.. there were tests.. special tests..*

A: *special tests..*

M: what comes here?.. '*varied ability*'.. [CP]

A: which? is it connected? [CP]

M: I don't know what.. '*varied ability*' is one word, no? [LAN]

A: varied ability.. *samazama na*..[TR]

M: varied ability [*noryoku*]

A: *special tests of..*

M: [INAUDIBLE]

A: I feel this is a verb..[WC]

M: then it's '*were*' [Gra]

A: I see..

M: there were.. it says there was something [*atta*], then it's explained later..[SSf]

A: *boxing and wrestling*.. too.. ah, I see.. and what's this special test? [CP]

M: '*of which*'.. what does it mean?.. that winner.. winner of.. I see.. in running.. something.. discus and javelin throwing and wrestling.. here four items are juxtaposed.. one two three four events.. [SPb/SPf]

A: what else.. running is a footrace, right? [CP] [KOW]

M: yeah, running is equivalent to footrace.. *discus* and *javelin* are here.. wrestling.. already here.. what else? [CP]

A: boxing!

M: *boxing*.. or field events.. *foot race*.. already here.. ah! horse racing isn't here [EP]

A: ah..

M: so.. a person who excelled in these events.. of which excelled.. so there must be boxing.. but wrestling is put at the end.. it seems they are listed in order of occurrence.. *javelin* before wrestling and so on.. de.. *boys*.. *horse-racing*.. let's put it for now.. we know this is some kind of event, right [UND]

A: yeah, let's write it

M: *horse-racing*.. or *boxing*.. whichever.. we don't know.. pentathlon is what you call the winner of the competition.. pentathlon [...] is the name of a person.. so pentathlon became the name or what you call the winner of the competition.. and then *varied ability*.. how does this connect to that..I don't understand the context.. [CP]

A: then what's here.. *tests*

M: [READS INAUDIBLY] what's this [...] tried? no.. let's skip it.. I don't understand it.. this here first.. it becomes a religious topic.. [LFN]

A: *the evening of the third day was*..

M: this must be '*the*'.. there isn't anything else we can think of.. no other possibility.. *the evening of*..

A: *spent*? *used*? [SOU?]

M: *used to*.. used for? [*tsukawareta*] offering? offer.. what's that? sacrificial offerings [*mitsugimono*].. offer .. devote [*sasageru*].. offer [*teikyou suru*].. heroes.. [READS INAUDIBLY] *the fourth day*.. *was set*.. something.. *the fourth day*.. [TR]

A: this is *night* [Diff]

M: must be *night*.. *night of the full moon was set*.. ah.. .

A: *night of the full moon was set*.. uhmm.. *was set officially*? holy day.. what's this? [CP]

M: sacred night and day [*shinsei na yoru-hi*] [TR]

A: so.. what happened on the fourth day? [CP]

M: *the fourth day*.. *night of the full moon*..

A: I can almost get it, but not.. hmm.. this [...] *on the* something *and last day*.. [UND]

M: this is also some listing in order..[LOG/EP?]

A: um.. third..

M: yeah, third and fourth, but nowhere does it say the game ends in four days [LOG/KOW]

A: does it say how many days somewhere else? [CP] [ca.8secs]

M: no, it's not written anywhere, so we can't assume the fourth day is the last day [EP/SP/LP]

A: let's just make something up.. something about sequence [*junban*]

M: yeah, something connected to sequence.. [L1Eq?]

A: yeah..

M: we know it's about sequence..

A: something *and last day* [READS INAUDIBLY] winner of '*the*' is the easiest

M: I think so too.. *the footrace*.. *footrace* appeared [a lot] before.. let's make it '*the*'.. there's no need **to** say '*important footrace*'.. *the winner of the footrace*.. ***the*** [EMPHATIC] *year of his victory*.. there's no adjective coming before '*year*' here [WC/GRA]

A: no, I can't think of any..

M: give.. his name is given to.. the commemorative year of his victory, right? [CP]

A: umm..

M: *so da yo ne* [that's right]

A: [LAUGHS]

M: what else could come here ? [CP]

A: there was [READS INAUDIBLY]..

M: what do you want to put? the central part is '*year*'.. something comes before.. some adjective [pertaining to] year.. article? what could it be? [GRA/WC]

A: '*the*' is fine! [INAUDIBLE] [???]

A, M: we're done? okay...

Appendix 6: Encouraging verbalisation

Encouraging verbalization

> "In other case studies in which English-speakers and Japanese-speakers were involved, participants seemed to be able to think aloud [more] easily in their mother tongue. However, the amount each student could verbalise was very diverse (i.e., some students were natural talkers, and some were not.) Thinking aloud in a 2nd language is very hard (and it is not easy even in a 1st language). Comparing 1st and 2nd language thinking-alouds seems to be very complicated."
> (Uzawa Kozue pers.comm.)

According to the Ericsson & Simon model of verbal reporting it is acceptable, even obligatory, to remind informants who fall silent to continue their monologue. This view is not universally accepted: Krings 1987:173 argues that it may be better to allow informants to pause at their own discretion, that researcher interruption may disrupt cognitive processes, and that think-aloud data is more valid when "minimal intervention on the part of the experimenter takes place, and no pressure to verbalize is exerted in any way."

The standard prompt is the instruction to 'Keep talking', and this is preferred on grounds of neutrality, as less likely to affect the content of the informant's think-aloud than a more specific prompt such as 'Tell me what you're thinking now.' All prompts and reminders are, except in those instances when informants fail to even notice them, to some extent interruptive. Informants do not always react well to prompts and may even be unable to accept that they had not in fact been verbalizing aloud at the time it was issued. The various means of 'prompting' informants to verbalise during their task processing, and these can be ranged along a cline of 'interruptiveness.' Prompts can also be categorized as 'random' or 'fixed' in terms of where or when they occur during processing. A third distinction, whose cognitively importance is not yet clear, lies in whether or prompts are 'aural' or 'written/graphic.'

Even if, as Ericsson & Simon's model claims (and Leow & Morgan-Short 2004 accept) verbal reporting does not significantly impact on how an individual processes the task in hand, it is plausible that this may become increasingly less true as the processing-cum-reporting task becomes more challenging. Interruption to the spontaneous flow of thought intuitively appears to be a factor that should (cf. the comments from Krings 1987, above) be minimised, yet the need to have informants keep up the flow of verbalisation means that reminders or prompts of some kind may be necessary. Prompt-forms which I have encountered in the verbal report literature, and/or trialled myself, include: individualised spoken prompts, spoken prompts to informants as a group, flashing lights or buzzers (individuals or group), individualised hand signals, and graphic cues on the task page.

"Keep talking"

The prompt used by many researchers is simply a spoken reminder to "Keep talking", issued whenever the informant has remained silent for *n* consecutive seconds. From my own observations, and my experience as a verbal report subject, this prompt seems to have minimal impact on task processing. Informants may or may not acknowledge the prompt, but even doing so does not appear to seriously interrupt their processing except in isolated cases.

Ironically, perhaps, the greater disruption may be to the researcher's own observation and recording. Even a straightforward and economical system of coding informant behaviours requires fairly constant attention, and the need to time silences can seriously impinge on other aspects of the researcher's workload. My own experience suggests that, because brief periods of informant silence are extremely common, longer silences are typically only even noticed once they have exceeded several seconds. A review of a single GL1 informant protocol reveals that the intervals between the previous informant verbalization and my reminder to 'Please keep talking' averaged approximately 12 seconds, and that the interval tended to become longer as the session progressed. Practice did not appear to make perfect.

Higher & lower technology

In a language laboratory group data-gathering session, I trialled the use of small orange 'attention-getting' lamps on the student consoles as prompts to informants to continue verbalising. My supposition was that this would perhaps be less disruptive to processing than an oral reminder. I had noted that informants would often respond verbally to oral reminders—sometimes with clear signs of irritation—and I was keen to obviate these exchanges. Just as was true of oral prompts, informants sometimes failed to notice the lamps flashing, but when they did notice the prompt they were less likely to acknowledge it verbally. (I confirmed in post-task interviews that informants were not sure that I was listening at the time the lamp signal occurred.) In a session in another (rather obsolescent) language laboratory, I found that when I used the teacher's microphone to orally prompt an informant others experienced an irritating buzzing sound in their headseats. My experiment of resorting to hand signal prompts was unhelpful in that it was hard for informants to know precisely who I was signalling to. In one-on-one sessions, however, I found that simply pointing to the microphone could adequately replace the standard oral prompt. Interestingly, any response this signal elicited from the informant almost invariably took a silent form, such a nod or thumbs-up gesture.

Graphic prompts

The methods above are all 'random' in that their timing is reactive and thus unpredictable. Another option is to place graphic prompts at intervals on the page (Cavalcanti 1987.) But where should these be located? The standard cloze task format provides its own reminder to verbalise in the form of (typically) a blank line every *n*th word. In think-aloud protocols, however, we find that the deletion itself is often 'passed over' with a silent pause or a kind of temporary 'place-holding' sound or word (*nantoka* is 'something'):

"the olympic games took place every *dühdüh / nantoka* years on the.."

which (according to the post-task comments of informants) appear to represent a means of maintaining the rhythm of the stretch of text in focus, and which in

turn often appear to be accompanied by a small head nod, finger tap or other physical gesture. At points other than a deletion, the informant may pause and verbalise about her processing, which means that the deletions themselves are not the only points at which processing takes place. In the extract below, for example, the (JL1) informant pauses between deletions (4) and (5) to reflect on possible fillers:

"...eventually *nantoka* its local character eventually something its local character [...] became first a *nantoka* event and after rules against foreign competitor.. had been waived.. so rules against foreign competitor been waived..ha.. international..[5 secs] so I guess it eventually lost its local character..."
(Jl1 consultant/informant A.)

It is widely held (Taft 1991) that sentence endings are major loci of text processing (although of course natural cloze deletions do not reliably occur at sentence ends) and the points at which the reader performs much of her cumulative 'sense-making' about a text. The ends of sentences thus create a kind of breathing space for readers to comment on the passage and provide something for them to say about it. In an effort both to minimise the need for spoken reminders, and to find out whether prompts at sentence ends might be valuable (and inspired by Cavalcanti op.cit.) I modified the cloze task format as shown below:

> In ancient Greece athletic festivals were very important and had strong religious associations. The Olympian athletic festival, held every (1)_____ years in honour of Zeus, eventually (2)_____ its local character, became first a (3)_____ event, and then, after the rules (4)_____ foreign competitors had been waived, international.•(5)_____ one knows exactly how far back (6)_____ Olympic Games go, but some official (7)_____ date from 776 B.C.•

A group of six JL1 informants (who had thought aloud in the regular way using the VIDEO RECORDER passage were asked to verbally report at every black dot (•) they encountered in the OLYMPICS passage as well as at any other points during their processing. The products of this 'dot task' format were compared with those of informants processing the same task without the sentence-end graphic prompts. The results were mixed: verbal reporting of 'dot task' processing proved not to be markedly greater in amount, but more of the reporting appeared to take place at sentence ends (i.e. at dot prompts) *at the expense of reporting at other stages*. These sentence end, dot-prompted verbalisations appeared seemed rather more structured or 'other-directed' (which feature partly inspired the NCR reporting format mentioned above.) As one might expect, 'dot' reporting featured rather less of the of the fragementary 'running verbalisation' found in most think aloud protocols, and in terms of the Ericsson & Simon model this must detract from its reliability and/or validity.

Extracts from 'dot' think aloud protocols are not shown here, as a very good picture of their content may be gained from the NCR protocol data cited in chapter 6. (NCR and 'dot' reporting clearly have a good deal in common, and both would perhaps be placed under the same heading (see chapter 4.15) in Cohen's taxonomy. In addition to 'dot' protocol data, insight into informants' response to the format were gleaned from a group discussion involving myself and as many (n=4) of the trial participants who could be brought together. One informant indicated that he had found it easier to balance the tasks of recovering deletions and "holding [herself] ready" to report on her processing because she knew that reporting would be required at predictable locations. The other informants participating in the discussion appeared to agree with this notion, and the distinction I drew at the head of this section between 'random' and 'fixed' prompts may be not only a psychologically real one, but also an important in terms of the think-aloud task. Here in table form are the main differences in verbal reporting noted between 'dot' and 'regular' task formats:

each other fairly circumspectly. Self-selection of pair partners is perhaps the best guarantee of this.

The Gricean maxim of economy was seen to apply to pair think aloud, however, in that (as I noted in a noted during a few post-task interviews) one partner would at times avoid telling the other something which she anticipated she would already know: the appearance of 'talking down' to a partner, post-task interviews suggested, was to be avoided. As noted in section 6.4, quite extensive use was made of 'face-saving' gambits in communicating information between partners, and it is hard to gauge how much of a problem this apparent withholding of pertinent cues or intuitions actually was, and indeed to tell whether these post-task mentions in fact pertain to cues that only become apparent to the individual post-task. I tried the tactic of emphasising to informants that they should try to "tell the tape" everything that seemed to be pertinent to a recovery. No 'omitted' information was mentioned by partners post-task when this reminder was given, but this is no guarantee that nothing was withheld. There seemed to be a strong element of chance in whether these items of previously unmentioned information arise at all in post-task interview, so it is difficult to judge how widespread the phenomenon actually is. On a positive note, we may assume that genuinely salient information will tend to be mentioned explicitly and/or made apparent in the filling of the item, so that the concern here is essentially with secondary or confirmatory cues

[3] With fixed, predictable reporting locations or occasions, informants' affective response to the think aloud task appears to be more positive; all participants who took both the regualar VIDEO RECORDER and 'dot' OLYMPICS tasks found the latter task less stressful and a majority opinion felt it was more rewarding in terms of their ability to report to their own satisfaction.

[4] The apparent consensus about 'holding oneself ready' to report suggests that think aloud informants may feel a certain tension between the recovery aspect of the task and that of reporting the recovery. In the words of one GL1 informant: "[You] want to have the feeling.. that you [i.e. the researcher] will more or less understand what gets said [in the think aloud]..otherwise there's no sense to it."

To sum up, even 'productive' think aloud informants have told me that they found the task tiring and often uncomfortable. On the other hand, the best informants may be those who have made a personal investment in carrying out the task. It seems reasonable (see chapter 8) therefore, to try to design verbal report tasks in such a way as to allow informants the maximum return on that investment, which can perhaps best be realized in a sense of satisfaction at having 'reported well'.

Information withholding in pair-condition reporting

Encouragement to verbalise may also be needed in pair-condition verbal reporting, but for affective/interactive reasons. Haastrup 1991 noted that in some instances one of the informants in a pair would fail to mention information pertinent to the lexical inferencing task but later reveal this in post-task interview. This behaviour, Haastrup concluded, was motivated by a fear of ridicule. In my own paired-informant data fear of ridicule did not appear to be a problem, as (with some exceptions in which the think aloud session had to be ended prematurely, or in a pairing like that of Frank & Anneke in which mutual teasing seemed to be accepted) informants appeared to treat

Appendix 7: VIDEO RECORDER passage and cloze

To produce a colour TV picture an enormous amount of information is required: on a TV the screen is scanned at 25 frames per second, which means that every second 25 separate pictures flash across the scereen of your TV. A one-hour recording alone consists of 90,000 separate pictures. Video recorders use magnetically coated tape just like a normal audio tape, only wider, inside a cassette. The tape travels fairly slowly, and as it travels the tape first passes an erase head which erases any previous signals, and then it travels around a fast-spinning drum. There are two video recording heads on opposite sides of this spinning drum, and the video track is recorded in diagonal stripes across the tape. Two striped contain the information required for a single picture, so that a three-hour video cassette, for instance, would have 540,000 stripes recorded on it.

After it leaves the spinning drum, the tape passes two more heads. These are the audio head and the control head. The audio head records the soundtrack along the top edge of the tape, and the control track which synchronises playback speed to recording speed is recorded along the bottom of the tape. Because the tape moves quite slowly, the sound quality of a video tape is not as good as that on a cassette recorder. Because of this, hi-fi videos have two extra audio tracks which are recorded in the stripes together with the video signals. To play back a video tape, the same process occurs in reverse, with the recording heads working as playback heads, and reading information instead of writing it.

280 words

FK 12

To produce a colour TV picture an enormous amount of information is required: on a TV the screen is scanned at 25 frames per second, which means that every second 25 separate pictures flash across the screen of your TV. A one-hour recording alone consists of

(1)..... separate pictures. Video recorders use a

(2)..... coated tape just like a normal (3).....

tape, only wider, inside a cassette. (4)..... tape

travels fairly slowly, and as (5)..... travels the tape

first passes an (6)..... head which erases any previous

signals (7)....., then it travels around a fast-

(8)..... drum which is carefully tilted at

(9)..... angle to the tape. There are (10).....

video recording heads on opposite sides (11).....this

spinning drum, and the video (12)..... is recorded in

diagonal stripes across (13)..... width of the tape. Two

of (14)..... stripes contain the information required for
(15)..... single picture, so that a three-
(16)..... video cassette, for instance, would have
(17)..... stripes recorded on it.

After it (18)..... the spinning drum, the tape passes
(19)..... more heads. These are the audio
(20)..... and the control head. The audio
(21)..... records the soundtrack along the top
(22).....of the tape, and the control (23)....., which synchronises playback speed to recording (24)....., is recorded along the bottom of (25)..... tape. Because the tape moves quite (26)....., the sound quality of a video (27)..... is not as good as that
(28)..... a cassette recorder. Because of this, (29).....
videos have two extra audio tracks (30)..... are recorded in the stripes together with the video signals.